Immunoinformatics of Placental Malaria Vaccine Development

Jessen, Leon Eyrich

Publication date:
2014

Document Version
Peer reviewed version

Immunoinformatics of Placental Malaria Vaccine Development

A DISSERTATION PRESENTED
BY
LEON EYRICH JESSEN
TO
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS, DEPARTMENT OF SYSTEMS BIOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
PHD - PHILOSOPHIAE DOCTOR
IN THE SUBJECT OF
BIOINFORMATICS / COMPUTATIONAL BIOLOGY

TECHNICAL UNIVERSITY OF DENMARK
KONGENS LYNGBY, DENMARK
APRIL 2014

Preface

About

1. Duration May 2010 - Apr 2014
2. Research Institution: The Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark.
3. Keywords: Bioinformatics, computational biology, statistics, immunology, vaccines, malaria, proteins/peptides, Genotype-phenotype correlation, high-density peptide microarray.

Collaborators

1. Professor Ali Salanti. Department of International Health, Immunology and Microbiology, Centre for Medical Parasitology, University of Copenhagen, Denmark. salanti@sund.ku.dk
2. Directeur, MD, PhD Philippe Deleron. Département Santé. Institut de Recherche pour le Développement (IRD). Faculté de pharmacie, Université Paris René Descartes - Paris 5, Paris, France. philippe.deloron@ird.fr
3. CEO Claus Schäfer Nielsen. Schafer-N. Copenhagen. Denmark. peptides@schafer-n.com
4. Professor Søren Buus. Department of International Health, Immunology and Microbiology. University of Copenhagen. Denmark. sbuus@sund.ku.dk

Supervisors

1. Professor Ole Lund (main). Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Denmark. lund@cbs.dtu.dk

2. Professor Lars Hviid (co). Department of International Health, Immunology and Microbiology, Centre for Medical Parasitology, University of Copenhagen, Denmark. lhviid@sund.ku.dk

3. Associate Professor Morten Nielsen (Unofficial co). Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Denmark and Instituto de Investigaciones Biotecnologicas, Universidad Nacional de San Martin, San Martin, B 1650 HMP, Buenos Aires, Argentina. mniel@cbs.dtu.dk

The dissertation at hand was prepared at the Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU) as partial fulfilment of the requirements for the degree of Philosophiae Doctor (PhD) in the subject of bioinformatics / computational biology.

This PhD programme was funded by: National Institutes of Health (NIH) [HHSN272201200010C]. EU FP7 PepChipOmics: The European Union 7th Framework Program FP7/2007-2013 [222773]. The Center for Genomic Epidemiology (www.genomicepidemiology.org) grant 09-067103/DSF from the Danish Council for Strategic Research. The University of Copenhagen - Program of Excellence: "Membrane topology and quaternary structure of key parasite proteins involved in Plasmodium falciparum malaria pathogenesis and immunity" and lastly the Life Sciences PhD programme at the Technical University of Denmark.

Leon Eyrich Jessen
April 2014

Please note that the section with supplementary materials, supporting this PhD thesis, is not included in the printed version of the thesis, but available at for download at http://www.cbs.dtu.dk/~jessen/phd_suppl_mat.pdf or at request by writing to jessen@cbs.dtu.dk The algebraic computations in this thesis were performed by using *Maple*™ *v17*. Molecular visualisations were done using the *PyMOL Molecular Graphics System, v1.5.0.4 Schrödinger, LLC*. Computational data analysis and depictions hereof, was performed using *R: A language and environment for statistical computing. R Core Team (2012). v2.15.2. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.*

Contents

Listing of figures

Til mormor og morfar.

Dissertation advisor: Prof. Ole Lund Leon Eyrich Jessen

Immunoinformatics of Placental Malaria Vaccine Development

Abstract

Malaria is an infectious disease caused by a protozoan parasite of the genus *Plasmodium*, which is transferred by female *Anopheles* mosquitos. WHO estimates that in 2012 there were 207 million cases of malaria, of which 627,000 were fatal. People living in malaria-endemic areas, gradually acquire immunity with multiple infections. Placental malaria (PM) is caused by *P. falciparum* sequestering in the placenta of pregnant women due to the presence of novel receptors in the placenta. An estimated 200,000 infants die a year as a result of PM. In 2004 the specific protein responsible for the pathogenesis of PM was identified as the *P. falciparum* Erythrocyte Membrane Protein 1 (PfEMP1) variant VAR2CSA. VAR2CSA is the leading candidate for a vaccine against PM. The thesis is divided into 4 parts, where part I provide the reader with an introduction and background for the subjects covered in the thesis. Part II presents the first paper: *"SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments"*. *SigniSite* is based on a non-parametric statistical evaluation of the positional distribution of amino acid residues in a multiple sequence alignment (MSA), thereby quantifying residue association to MSA phenotype. *SigniSite* was found to outperform comparable state-of-the-art methods. Furthermore part II addresses the issue of controlling type I and type II error probabilities in multiple testing scenarios and lastly the

analysis of the MHCI:peptide binding interaction by application of the *SigniSite* method. Part III presents the second paper: *"Insight into Antigenic Diversity of VAR2CSA-DBL5ε Domain from Multiple Plasmodium falciparum Placental Isolates"*. The data consisted of 70 VAR2CSA-DBL5ε sequences each with associated phenotypes. Immunity towards PM is gradually acquired, therefore if a given sequence motif can be phenotype-correlated then the motif may be involved in VAR2CSA immunogenecity. Motifs defining VAR2CSA immunogenecity are naturally interesting in vaccine development context. The motif 'TFKNI' was found to be correlated with the birth weight of the child. Part IV presents the development of two methods for analysis of high-throughput data from a novel *High Density Peptide Microarray* (HDPMa) chip technology. Subsequently the HDPMa chip is applied for the discovery of linear B-cell VAR2CSA epitopes. Peptides 'GMDEFKNTFKNIKE' and 'SCGSARTMKRGYKNDNYELCKYC' were identified as linear B-cell epitopes. The latter subsequently experimentally found to be highly immunogenic, but not capable of blocking VAR2CSA:receptor interaction. In summary, the work described in this thesis centres around the development and application of bioinformatics tools for *in silico* analysis of VAR2CSA, with an emphasis on statistical methodology. It is the hope of the author that the tools, developed, presented and applied in this thesis, may serve as an offset for further research and development in the field of placental malaria vaccine development.

Immunoinformatics of Placental Malaria Vaccine Development

Resume

Malaria er en infektionssygdom, forårsaget af en protozoisk parasit af slægten *Plasmodium*. Parasitten overføres af hunmyg af arten *Anopheles*. WHO estimerer at der var 207 millioner tilfælde af malaria i 2012, hvoraf 627.000 var fatale. Mennesker, som bor i malariaendemiske områder, erhverver gradvist immunitet, med stigende antal infektioner. Placental malaria (PM) forårsages af *P. falciparum*, som afsondres i moderkagen hos gravide kvinder. Dette skyldes tilstedeværelsen af nye receptorer i moderkagen. Det estimeres at der hvert år dør 200.000 spædbørn af PM. I 2004 blev det specifikke protein, som er ansvarlig for patogenesen af PM, identificeret som *P. falciparum* Erythrocyt Membran Protein 1 (PfEMP1) varianten VAR2CSA. VAR2CSA er den førende kandidat til en vaccine mod PM. Denne ph.d. afhandling er inddelt i 4 hoveddele, hvor del I forsyner læseren med en introduktion til og baggrund for emnerne, indeholdt i afhandlingen. Del II præsenterer den første artikel: *"SigniSite: Identifikation af aminosyrerest-niveau genotype-fænotype korrelationer i multiple sekvens alignment af proteiner"*. *SigniSite* er baseret på en ikke-parametrisk statistisk evaluering af den positionelle aminosyrerestfordeling i et multipelt sekvensalignment (MSA), hvorved aminosyrens associering med MSA fænotypen kvantificeres. Vi fandt at *SigniSite* var i stand til at performe bedre end den nyeste sammenlignelige metode. Ydermere adresserer del II udfordringen med at kontrollere

sandsynligheden for fejl af type I og type II i multiple testscenarier. Sluttelig anvendes *SigniSite* til at analysere MHCI:peptide bindingsinteraktionerne. Del III præsenterer den anden artikel: *"Indsigt i den antigeniske diversitet af VAR2CSA-DBLε domænet fra multiple Plasmodium falciparum placentale isolater"*. Dataene bestod af 70 VAR2CSA-DBL5ε sekvenser, hver især associeret med et sæt af fænotyper. Immunitet mod PM erhverves gradvist, hvorfor såfremt et givet sekvensmotiv kan fænotype korreleres, så er dette motiv muligvis involveret i VAR2CSA immunogenicitet. Motiver, som definerer VAR2CSA immunogenecitet er naturligvis interessant i kontekst med vaccineudvikling. Motivet '`TFKNI`' korrelerede med barnets fødselsvægt. Del IV præsenterer udviklingen og anvendelsen af to metoder til analyse af høj-produktionsdata fra en ny *Høj Densitets Peptid Mikroarray* (HDPMa) chip teknologi. HDPMa chippen blev efterfølgende anvendt til opsporing af lineære b-celle epitoper. Peptiderne '`GMDEFKNTFKNIKE`' og '`SCGSARTMKRGYKNDNYELCKYC`' blev identificeret som lineære b-celle epitoper. Den sidste af de to, blev efterfølgende eksperimentelt fundet til at være høj-immunogent, men ikke i stand til at blokere VAR2CSA:receptor interaktionen. I sammendrag, centrerer arbejdet, som beskrives i denne ph.d. afhandling sig om udviklingen og anvendelsen af bioinformatiske værktøjer til *in silico* analyse af VAR2CSA, med særlig vægt på statistisk metodologi. Det er forfatterens håb at værktøjerne udviklet og anvendt i denne ph.d. afhandling, kan tjene som udgangspunkt for, eller blot bidrage til, videre forskning og udvikling i feltet for placental malaria vaccineudvikling.

Acknowledgments

I have been fortunate enough to be a part of The Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark since September 2009, first as a master student and subsequently as a PhD student.

There is no questioning that CBS truly is a unique place. The publication track record is a testimony to the high academic level. By populating CBS with an academic staff of great diversity, consisting of engineers, medical doctors, physicists, molecular biologists, mathematicians, computer scientists, chemists, etc., Søren Brunak has established a bioinformatics centre fully capable of competing in the top of international academia. However not only is the academic level very high, but additionally the people at CBS are friendly and forthcoming and despite a high work pace and busy schedules the occasional 5-10 minute talk is never further away then the two *'La Cimbali'* italian coffee makers in kitchen. Passing remark - I did not really drink coffee until I started at CBS, that has changed now! I think CBS can be summed in how the people at CBS refer to themselves: *You're not simply an employee - You're a CBSian!*

With a background, which can best be described as wet-lab molecular biology, I entered unchartered territory and met a world populated with people speaking in strange tongues, hearing words such as: *linux, gawk, perl, python, R, PyMOL scripting, and parsing* on a daily basis. I did not speak the language of the natives - But as with the coffee consumption, that has changed! I wish to thank everyone

at CBS for contributing to the special atmosphere and for making CBS such a nice place to be!

I have had the pleasure of having professor Ole Lund as a supervisor. Ole is, amongst other things, Vice Center director at CBS, group leader of the *Protein and Immune Systems Biology* group at CBS, and head of the PhD committee. Despite all of this Ole is always willing to help and give input on ongoing projects, discuss ideas etc. and only seldom is it not possible to book a meeting with Ole from day to day. Often meeting can be arranged simply by dropping by his office. Ole facilitates the settings for completing a PhD programme at not only a high academic level, but also in a group where social interaction is weighted as an important part of obtaining your PhD! Thank you very much Ole for being a great supervisor and for presenting me with this unique academic opportunity under your very competent supervision!

Associate professor Morten Nielsen has played a vital role in me getting my PhD. Morten has been kind enough to function as an un-official co-supervisor, providing invaluable feedback. Morten is like Ole a very busy man, but always willing to help and discuss, be that face to face or via Skype to Buenos Aires. Thank you Morten for helping out, whenever it was needed!

A special thanks to out primary collaborator professor Ali Salanti from the Centre for Medical Parasitology, University of Copenhagen for numerous interesting meetings and for providing me with data to work on and for taking my results from dry-lab to wet-lab, despite the varying success - Thanks Ali!

Many thanks are also due to Lars Hviid for helping making this PhD programme possible!

The vast majority of my time at CBS, I have been in office 003, which is a very special office! I have had the privilege of sharing office with such great people as: Massimo, Bent, Edita and Simon. I am sorry guys, but a special thanks must go to Edita for always being able to spot "when I needed a piece of chocolate"! My lunch-buddy 'Burger-Bent' however also deserve a special thanks for making sure, that I was never low on electrolytes!

Naturally thanks also go to the research group formally known as the

Immunological Bioinformatics group, which recently has changed name to *Protein and Immune Systems Biology*. You are a great bunch of people and I have enjoyed not only the academic discussions at our weekly monday, but also our group-nights out on town!

Further special thanks go to Katrine Juul for proof-reading the math notation, impressively noting the appearance of the mountains! My mother and my mother-in-law for helping with out with the rugrats, while I have been busy writing this thesis.

Last, but not least I am lucky beyond any comprehension to share my life with my wonderful wife Britta, without whom none of this would have been possible! Also a very special and loving thanks to our two incredible boys Gustav and Viktor, for being such a couple of wonderful rugrats! The life I have with you is so much more than any man could ever hope for! I honestly humbly thank you for your endless support, your love especially in the final stages of completing my thesis, where Britta managed to squeeze a final proof-reading! Thank you so much!

THANK YOU ALL!

Papers Included in this Thesis

- Paper I:
  Insight into antigenic diversity of VAR2CSA-DBL5ε domain from multiple Plasmodium falciparum placental isolates.
  Gnidehou S, **Jessen L**, Gangnard S, Ermont C, Triqui C, Quiviger M, Guitard J, Lund O, Deloron P, Ndam NT.
  PLoS One. 2010 Oct 1;5(10). pii: e13105. doi: 10.1371/journal.pone.0013105. PMID: 20957045, PMCID: PMC2948511

- Paper II:
  ***SigniSite*: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments.**
  Jessen LE, Hoof I, Lund O, Nielsen M.
  Nucleic Acids Res. 2013 Jul;41(Web Server issue):W286-91. doi: 10.1093/nar/gkt497. Epub 2013 Jun 12. PMID: 23761454, PMCID: PMC3692133

Abbreviations

A	AMP	Anti Microbial peptide
	APC	Antigen Presenting Cell
	AUC	Area Under the Curve
B	BCR	B-Cell Receptor
C	CAS	Computer Algebra System
	CBS	Center for Biological Sequence Analysis
	CDF	Cumulative Distribution Function
	CMP	Centre for Medical Parasitology
	CMT	Correction for Multiple Testing
	CSA	Chondroitin sulfate A
	CSPG	Chondroitin sulfate proteoglycan
	CTL	Cytotoxic T-Cell
D	DBL	Duffy binding-like
	DiCo	Diversity Covering
	DMD	Digital Micro mirror Device
	DNA	Deoxyribonucleic acid
	DTU	Technical University of Denmark
E	ECTS	European Credit Transfer and Accumulation System
H	HD	High Definition
	HDPMa	High Density Peptide Microarray
	HIV	Human Immunodeficiency Virus
	HIVdb	The Stanford University HIV drug resistance database
	HLA	Human Leukocyte Antigen
I	IE	Infected Erythrocyte
	ILC	Innate lymphoid cell
	Ig	Immunoglobulin
K	KU	University of Copenhagen

M	MHC	Major Histocompability Complex
	MSA	Multiple Sequence Alignment
P	PAM	Pregnancy Associated Malaria
	PCC	Pearson's Correlation Coefficient
	PDB	Protein Data Bank
	PfEMP	*Plasmodium falciparum* Erythrocyte Membrane Protein
	PM	Placental Malaria
R	RBC	Red Blood Cell
	RGB	Red-Green-Blue
	ROC	Receiver operating characteristic
S	S/N	Signal-to-Noise
	SCC	Spearman's Correlation Coefficient
	SSN	Standard Score Normalise
T	TAP	Transporter Associated with Antigen Processing
	TCR	T-Cell Receptor
W	WHO	World Health Organisation

Part I

Introduction and Background

Life begins at the end of your comfort zone.

Neale Donald Walsch

1 Introduction

March 1st 2010 I handed in my thesis for the degree of master of science in engineering. Here I sit once again, trying to persuade LaTeX to yield the expected output, finding the proper references to build the introduction, which is to serve as the foundation for the entire thesis. In the midst of all this, I realise one important thing. Except for meetings and conferences - For the past 4 years I have been sitting at the same desk day after day, analysing, thinking, deliberately challenging myself and constantly ending up in situations, where I had no clue about the answer or how to get there or even how to begin. Physically I have not moved for the past 4 years, but I now have a much more profound understanding of the quote in the upper left corner of this page and have thusly arrived at my personal mantra: *I'll figure it out - I always do!*

1.1 PhD Project Overview

1.1.1 PhD Partners

This PhD project was initiated as a collaboration between the Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU) and the Centre for Medical Parasitology (CMP), Department of International Health, Immunology and Microbiology, University of Copenhagen (KU) and the PepChipOmics project, 7th Framework Programme for Research and Technological Development (FP7) of the EU, HEALTH-2007-1.1-4.

1.1.2 The CMP VAR2CSA Team

The VAR2CSA research team at CMP, lead by Professor Salanti, is world leading in the search for a placental malaria vaccine. In 2004 they identified the PfEMP1 variant VAR2CSA as the protein responsible for placental malaria [129], in 2010 they managed to recombinantly produce full-length VAR2CSA [83] and in 2012 the team received substantial financial support for preclinical development of a VAR2CSA-based vaccine from the EU Eurostars program.

1.1.3 PhD Project Objective

The objective of this PhD programme was: *"The development and application of computational tools aiming at obtaining a better understanding of how VAR2CSA sequence variation affects immunogenicity and capability to induce parasite adhesion blocking antibodies. The ultimate goal being able to produce a vaccine, which can be used to protect pregnant women against Placental Malaria."*

1.1.4 Additional projects not included in this PhD thesis

1. Development and Application of VAR2CSA Diversity Covering Sequence Generator aiming at creating a set of diversity covering sequences for

inducing broad immune response towards VAR2CSA.

2. Analysis of sequence data from Malaria Protein domain DBL3X. External collaborators: Centre for Medical Parasitology, University of Copenhagen, Denmark.

3. Prediction of HIV-1 protease inhibitor drug resistance. External collaborators: Department of Biology, Biomolecular Sciences, University of Copenhagen, Denmark.

4. NGS: Next Generation Sequencing - Full genome sequencing and assembly of Plasmodium falciparum genomes, aiming at elucidating genomic stability. External collaborators: Center for Genomic Epidemiology, Technical University of Denmark, Denmark.

1.2 Immunology

1.2.1 Introduction to the human immune system

To put in an popular tongue, there is a constant arms race going on inside every human being. Pathogens constantly invade our body seeking nutrition aiming at multiplying in order to preserve their genetic material. As a counter measure, the immune system is constantly monitoring the health of every cell in out body along with the composition of extracellular fluids and tissues. The mechanisms by which the body is monitoring itself are highly efficient. In fact so efficient, that it has been proposed to mimic the system when monitoring IT infrastructure [34]. Which is an interesting extrapolation if you are a computational biologist, with a background in engineering.

The innate and the adaptive immune systems make up the human immune system [30, 31]. Traditionally has been viewed as two distinct mechanisms, new research however suggests that innate and adaptive immunity to a much higher degree are intertwined [92, 93]. In the following, the traditional view will be described, it should however be noted that there still is much to learn about the

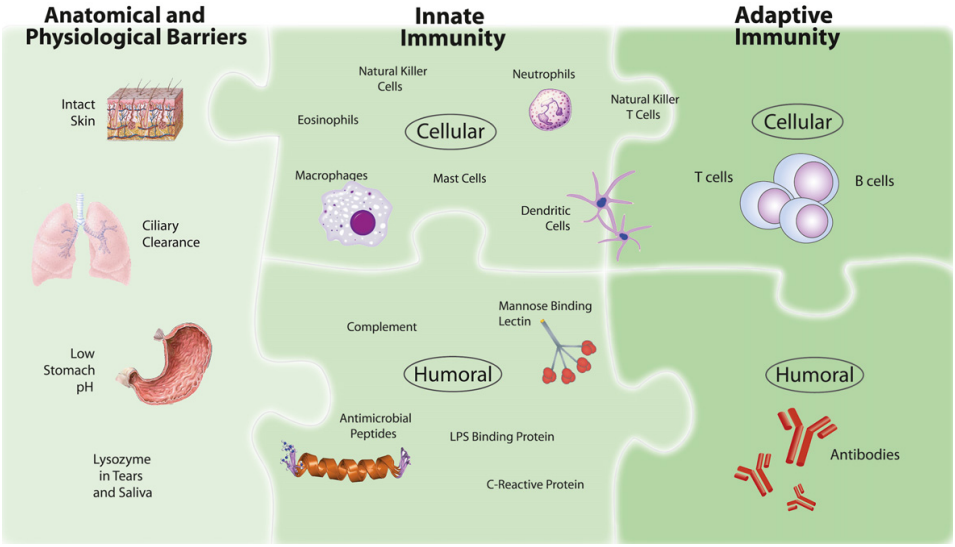


Figure 1.2.1: Simplistic 3-level view of the human immune system. Initially microbial infection is prevented by physical and chemical barriers, such as the skin and the low stomach pH. Should these fail, the immune system can mount a swift, but non-specific response via innate immunity. Adaptive immunity is responsible for long term immunity following an infection [147].

immune system and that the knowledge about it is highly dynamic. E.g. recently a new population of immune cells known as Innate lymphoid cells (ILCs) have been discovered [153]. These ILCs cannot be classified using the traditional system, due to lack of specific surface markers [141].

Simplified, the immune system consists of 3 levels: *i.* Physicochemical barriers, *ii.* The innate and the *iii.* adaptive immunesystem (see figure 1.2.1).

The innate immune system, also known as the non-specific immune system, provides an immediate response to infection, but does not give rise to long term immunity [147]. Innate effectors include e.g. anti-microbial peptides (AMPs) [55, 134], which are present on our skin and intestinal tract [53] protecting against invasion of opportunistic pathogens [65], such as e.g. *Staphylococci*, which are abundant [113]. I.e. non-specific effectors targeting common pathogenic structures. The innate immune system is considered ancient and shared across a wide variety of species [35]. The adaptive immune system on the other hand is

unique to vertebrates and is responsible for long term immunity [22]. The adaptive immune system is capable of recognising and clearing highly specific non-self structures, but initial response is slow [22]. If a previously recognised non-self structure is encountered, the response will however be massive and swift [104]. This strong and specific response elicited when encountering a previously 'seen' pathogen structure is the reason that he activation of the adaptive immune system is a paramount step in vaccine development. Long term immunity can be achieved via two types of responses: Cell-mediated and humoral.

1.2.2 Cell mediated immunity

Some microbes such as vira or certain types of parasites, will invade human cells in order to use the cell as a factory for proliferation. These intra-cellular pathogens are not directly visible to the human immune system, but rather indirectly via antigen presentation using the Major Histocompability Complex (MHC) class I partway [158] (See figure 1.2.2).

On the surface of every nucleated cells in the human body, sits the MHC class 1 molecule. The purpose of the MHC class 1 is to reflect the inner protein composition, by presenting samples of the current protein pool inside the cell.

Inside every living cell, there is a constant production and degradation of proteins (protein turnover). Various proteins are constantly synthesised and released into the cytoplasm [99]. These cytoplasmic proteins are constantly being degraded by the proteasome into free amino acids and smaller chains of amino acid residues referred to as peptides [94]. The free amino acids are recycled for new protein synthesis, whereas the peptides enter the endoplasmic reticulum through a system known as 'the transporter associated with antigen processing' (TAP). Once inside the endoplasmic reticulum, helper proteins called chaperones assist the pairing of peptides, typically 9-mers, and MHC class 1 molecules. The formed MHCI:peptide complex is transported through the Golgi Apparatus to the surface of the cell, where it is made visible to the immune system. [158]

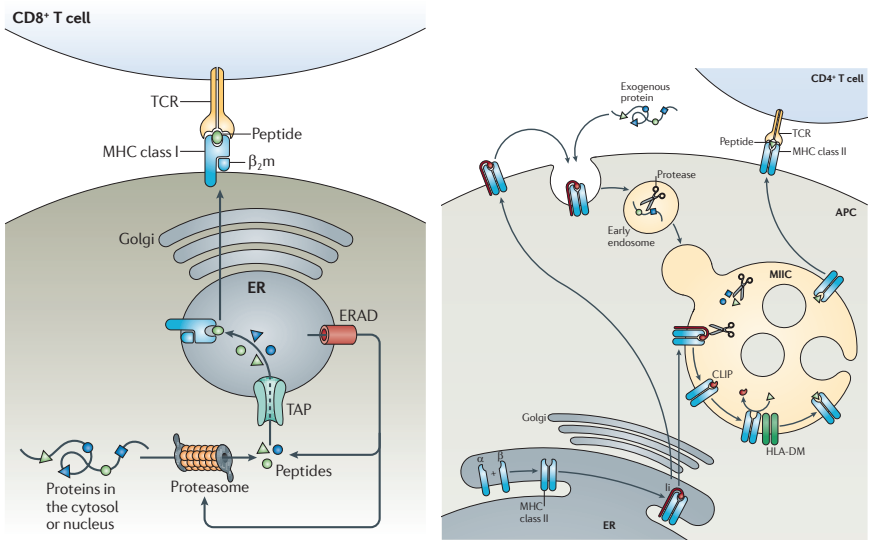


Figure 1.2.2: Antigen presentation through the MHC class I and II pathways [109]. Details in text.

The CD8+ cytotoxic t-cells (CTLs) are constantly monitoring the health of each cell, by coupling of the t-cell receptor (TCR) to the presented MHCI:peptide complex [114]. Each CTL recognises one specific target. If a MHCI:peptide complex is recognised, the CTL will adhere through the MHCI:peptide:TCR interaction. The interaction is further stabilised by the CD8 co-receptor (hence CD8+). This is the first signal in a tightly controlled two step activation [20]. The second signal comes from the stimulation of the CD28 receptor on the CTL by either the CD80 or the CD86 molecules on the infected cell [36]. The exact mechanism of t-cell activation has yet to be eluded [14]. The result of this two-step activation of the CTL, is clonal expansion, resulting in a dramatic increase in CTLs specific for the pathogen derived peptide, presented on the surface of the infected cell. Each of these CTLs are activated and capable of inducing programmed cell death (apoptosis). Figure 1.2.2 illustrates the process.

During apoptosis, the cell undergo a series of changes resulting in the complete, yet controlled disintegration of the cell into small packages, which will

be taken up by macrophages. The macrophages serve as bins for the immune system, devouring anything tagged for destruction. It is obviously paramount that this system is tightly controlled.

1.2.3 Humoral immunity

B-cells, a specific type of immune cell, is found throughout the human body. Immobilised on the surface of the b-cell, sits up to 120,000 b-cell receptors (BCRs) [156]. The specificity of each BCR is extremely high and the repertoire of unique BCRs is potentially up to 10^{11} [60] in each human. As with the TCR, BCRs will recognise non-self structures with a very high specificity, e.g. a surface protein of a bacterium. Once a b-cell encounters and binds such a non-self structure, the BCR:antigen binding trickers a signal cascade leading to the uptake and subsequent degradation of the antigen. It has been shown, that the B-cells are capable of extracting pathogen antigens tightly anchored in a non?internalisable surface [12]. This way, the B-cell is capable of 'stealing' the antigen from the surface of e.g. a bacterium and process it as follows: The BCR bound structure is transported to the Golgi apparatus, where it is coupled with a major histocompability complex class II (MHCII) molecule [13]. The MHCII:peptide complex is then transported to the surface, where it is presented. The B-cell have thus become an antigen presenting cell (APC). MHCII:peptide complex can be recognised by the TCR of a T-helper cell (CD4+) [116] (See figure 1.2.2). The T_h trickers full activation of the b-cell. The b-cell will proliferate and divide into effector b-cells and memory b-cells. Some of the effector b-cells will become plasma b-cells capable of releasing vast amounts of soluble BCRs. Soluble BCRs are commonly known as antibodies. [116]. Each released antibody has the same specificity as the BCR which initially recognised the pathogen structure, the antibodies will following release, target these pathogen structures and tag them for pick up by macrophages. This process is known as opsonisation. Furthermore the antibody tag of the pathogen can lead to inhibition of pathogen surface antigens, which are essential for pathogenesis.

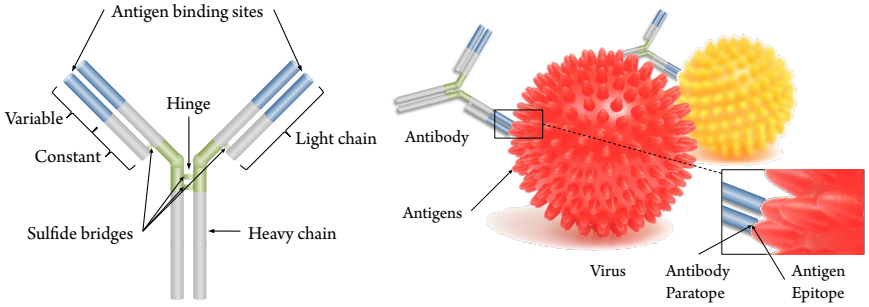


Figure 1.2.3: Simplistic model of an antibody and antibody-antigen interaction. Antibody model adapted from [15].

1.2.4 Antibodies, epitopes and paratopes

A key component of the humoral response from the adaptive immune system is the antibodies [147]. Antibodies are proteins capable of highly specific interaction with other molecules. An antibody consists of two light chains and two heavy chains, held together by sulfide bridges. Both the light and heavy chain have a constant and a variable part. The variable part determines the specificity of the antibody and the antigen binding part is located distal from the hinge region connecting the two heavy chains [15] (See figure 1.2.3).

Five type of human antibodies exist: immunoglobulin (Ig)G, IgA, IgM, IgE and IgD [108]. Antigen-binding non-immobilised antibodies in the bloodstream of the IgG type are important for protection against pathogen infections [136], whereas antigen-binding immobilised antibodies, found on the surface of b-cells, are of the IgD and IgM types [108].

The amino acid residues on the antigen binding sites, which interact directly with the antigen, constitute the paratope. The amino acid residues on antigen, which interact directly with the paratope, constitute the epitope [151]. Monoclonal antibodies target the same epitope on the same antigen, whereas polyclonal antibodies target different epitopes on the same antigen.

Epitopes can be either continuous (linear) or discontinuous (conformational) [151]. A linear epitope, consists of a continuous stretch of amino acid residues,

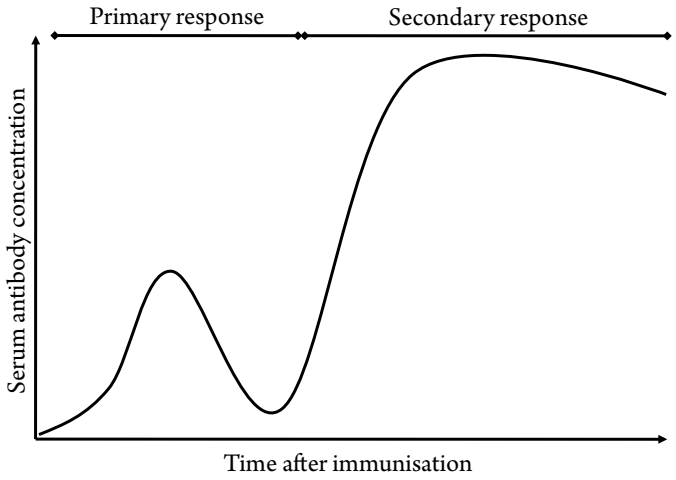


Figure 1.2.4: Primary and secondary immune response. The primary response takes time to develop. Upon second challenge with the same antigen, the secondary response is much stronger and faster. Adapted from [84].

e.g. a protruding loop. Conformational epitopes, on the other hand, are made up of amino acid residues brought together by the tertiary structure of the protein. The majority of epitopes are discontinuous [138]. But the vast majority harbours a linear determinant (>5 residues) [66, 87].

The identification of which structures are responsible for the elicited immune response represents the initial and crucial step of vaccine development.

1.2.5 Vaccine strategies and development

Briefly on vaccines

A vaccine consists of a biologically active structure, capable of activating the immune system in such a way, that a future exposure to the actual pathogen will lead to a swift clearance, but without any of the symptoms of infection. Fig. 1.2.4 illustrates the primary and secondary immune response, when encountering the same antigen twice. The first response takes time to develop and at peak is not as strong as the secondary response. Furthermore upon second encounter, the

response is swift.

The purpose of a vaccine is to induce the primary response in a harmless way, such that upon infection with the pathogen, from which the vaccine has been derived, the response will be quick and the pathogen will be cleared.

On the Origin of Vaccines

Vaccines represent one of mankind major health related scientific breakthrough and have virtually eliminated a wide range of otherwise potentially fatal childhood diseases [120]. Vaccines are routinely administered in the developed world. In the US approximately 80 - 90% of all infants are vaccinated against diphtheria, pertussis, tetanus, measles, mumps, rubella, polio, Haemophilus influenzae type b, hepatitis B, varicella, pneumococcus and hepatitis A [149].

Edward Jenner (1749-1823), an English physician, made the discovery laying the foundation for the later smallpox vaccine and the subsequent eradication of smallpox [18, 143]. Jenner inoculated a boy with cowpox and subsequently exposed the boy for smallpox, the result of which was that the boy was protected [123]. The word vaccine originates from the latin 'vacca' meaning 'cow' [8]. Figure 1.2.5 depicts Jenner as he is vaccinating a baby [139]. Continuing Jenner's work, Louis Pasteur (1822-1895), Paul Ehrlich (1854-1915) and Emil Behring (1854-1917) became known as 'The fathers of immunology and vaccinology' [81]. The variola virus, the cause of smallpox, was completely eradicated in 1977 [139].

Vaccine Strategy

An efficient vaccine must activate an immune response, as good as possible compared to that, which would be elicited by an actual infection, without inducing disease symptoms. A major challenge in vaccine development is therefore that the efficiency of the vaccine is proportional to the immunogenecity, but as the immunogenecity increases, so does safety and

Figure 1.2.5: Engraving depicting Edward Jenner (1749-1823), while he vaccinates a baby against smallpox. Jenner is considered the father of vaccinations, as he realised that human immunisation with cowpox, provided protection from smallpox [139]. ©2002 Science Photo Library.

tolerance issues [8]. One approach for mimicking the immune response from an actual infection is by using a live attenuated vaccine. A live attenuated vaccine is produced by repeated cultivation of the pathogen until it no longer retains pathogenicity. As pathogenicity is associated with an energy expense, e.g. the production of a specific toxin or a protein based pumping system, etc. it is a disadvantage, when growing under conditions, where pathogenicity is not a requirement for survival. Repeated rounds of cultivation, will therefore result in the non-pathogenic variant outcompeting the pathogen variant. A similar outcome can be obtained by using radiation [98]. Is has been argued that using an attenuated approach for malaria vaccination, may protect as much as 90% [74]. There are however several drawbacks using the live attenuated vaccine system. The primary one being the risk of reverting from non-pathogenic to pathogenic.

An example of vaccine reversion is the case, where the pathogenicity is located on a small extra-chromosomal circular piece of DNA, a plasmid, which can be transferred between organisms and thus result in what is known as horizontal gene transfer. In cases where the lack of pathogenicity is a result of a single mutation, the inverted mutation may result in a regain of pathogenicity. In the case of viruses, the combination of two non-pathogenic variants, may result in reformed virulence [95]. Once inside the human body again, the reason for the

original development of pathogenicity, i.e. the ability to proliferate by immune evasion, once againg represent a prerequisite for survival. Furthermore many pathogens are notoriously difficult to culture *in vitro* [118] making vaccine production complicated and therefore costly.

Since reversions poses a great risk, another approach is to identify the exact structure responsible for the elicited response. Moreover one may even identify the exact substructure and if feasible one can produce a so called subunit vaccine. By identifying the exact epitope(s), one may identify the "minimal immunogenic region of a protein antigen" [118]. Smaller structure may often be produced recombinantly and thus offers a cheap way of mass production. This minimum-region approach however poses the challenge of covering the diversity in cases where the immunogenic region displays a high degree of polymorphism. Furthermore when aiming at a cellular response, where epitopes must be matched with the specific variant of the MHC complex (the human leukocyte antigen (HLA) haplotype), the challenge is to make a vaccine, which cover the entire HLA haplotype diversity within the population. This complication arises from the fact that different populations carry different HLA haplotypes and since the haplotype:peptide interaction is unique, a peptide inducing an adequate response in one individual, may have no effect in another, see [118] for review. Lastly subunit vaccine may have a high efficacy, when looking only at the haplotype::peptide interaction, but when challenged with the complete pathogen, the subunit vaccine, may direct the response against parts, which are not targets.

1.3 Malaria

1.3.1 Historic Background

Malaria as a word, was first used in scientific context in 1827 by John MacCulloch [67]. The word malaria is derived from italian meaning 'bad air' [67, 101], recognising that the symptom could be observed in people living in damp and

swampy areas. Malaria-like symptoms is described in ancient Chinese writings dated approximately 2700 BC and later in writings of the greek philosopher Hippocrates from approximately 400 BC [97]. Using ancient DNA as evidence, it has been argued, that Malaria had a significant impact on mortality rates in ancient roman Italy [131]. Up until 1860 malaria posed a significant health risk to people living in the Lambeth and Westminster areas of London, only after drainage the risk was markedly reduced [101]. In summary, it is evident, that malaria has been with us for a very long time and more importantly, in the context of vaccine development, that Malaria has co-evolved with humans.

1.3.2 Malaria in the modern world

Today malaria affects the world on a global scale. In the 2013 World Malaria Report, WHO estimates that in 2012 there were approximately 207 million malaria cases, 627 000 malaria deaths [155] and that half the world population at risk of being infected with malaria [155]. The majority of infections occur in infants and small children under the age of 5 [155]. Parents are forced to stay home and take care of the children, the consequence of which is a markedly labour force reduction in malaria endemic countries [128]. This way malaria is not only affecting people on a personal level, but to a high degree also affecting the economy of the developing world [154]. Fig. 1.3.1 show the worldwide distribution of malaria along with the areas contribution to the total death toll [3]. From this, it is clear that sub-saharan by far bears the heaviest burden, when it comes to malaria.

1.3.3 Malaria Biology

Human malaria is caused by a small protozoan parasite of the genus *Plasmodium* [38]. 4 different species capable of human infection exists, namely: *P. falciparum*, *P. vivax*, *P. ovale* and *P. malariae* [122]. The closest known non-human malaria species is *P. reichenowi*, which causes mild malaria in chimpanzees [117]. *P.*

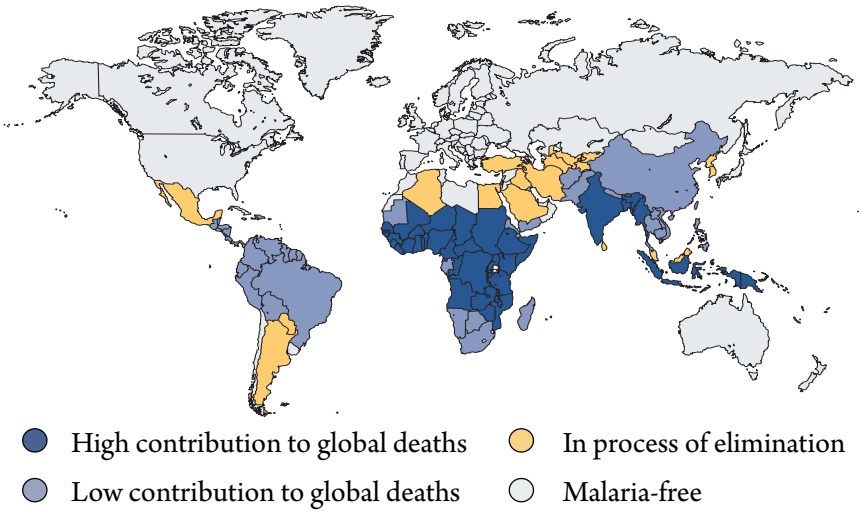


Figure 1.3.1: Worldwide distribution of malaria and contribution to death toll. Adapted from [3]

reichenowi / *P. falciparum* divergence is thought to have occurred 6 to 10 million years ago [46].

In 2009 a new *Plasmodium* species, isolated from two Chimpanzees living in close contact with humans, capable of both Chimpanzee and human infection, was identified [112]. Also *P. knowlesi* has been isolated from humans living in close contact with macaques [42]. Furthermore *Plasmodium* lineages have been identified in Gorillas along with *P. falciparum* [117]. The fact that several *Plasmodium* species have been found capable of host switching emphasises the capability of malaria to adjust.

A specific subtype of malaria is 'pregnancy associated malaria' (PAM) or placental malaria, which is caused by the *P. falciparum* species [50]. As the overall focus of this PhD programme has been on placental malaria, the focus will henceforth be on *P. falciparum* malaria.

The life cycle of *Plasmodium falciparum*

The lifecycle of the malaria parasite is complex and involves several distinct stages in both the human host and the *Anopheles* mosquito transmission vector [37].

The sporozoite form of the malaria parasite is found in the salivary glands of the mosquito [38] and infection occurs just after the mosquito bite [105]. During the bite, the mosquito injects saliva containing anticoagulants in the blood, preventing the blood from clotting [126]. Upon infection the parasites will seek out the liver and enter the hepatocytes, where it will generate thousands of merozoites [102]. The hepatocyte eventually ruptures releasing the merozoites into the blood stream, each of which is capable of infecting an erythrocyte [38]. After 48 hours the infected erythrocyte (IE) ruptures [63] and releases 16-32 daughter merozoites into the blood stream[102] and the cycle continues. This cycle is the reason for the 48hour fever attacks malaria gives rise to [63]. Inside the erythrocyte some of the parasites will form male and female gametocytes, which can then be taken up of another mosquito [90]. This way, the parasite has a asexual stage in the host and a sexual stage in the vector [121], the regulation of gametocytogenesis is not fully understood [43]. Figure 1.3.2 depicts this lifecycle.

Plasmodium falciparum immunity

The human immune system detects extracellular infections by activating systems, which labels and neutralises the pathogen [96]. Intracellular infections of nucleated cells are detected via surface presentation of pathogen antigens [91] (See sec 1.2). Upon malaria infection, an immune response will be elicited directed at different stages of the malaria life cycle [27]. The parasite has evolved a series of mechanisms of immune evasion [142]. By invading erythrocytes, the parasite exploits the erythrocytes lack of a nucleus, thereby circumventing the primary intracellular pathogen detection system [72]. As erythrocytes are non-nucleated cells, they lack the ability to present peptides derived from intracellular proteins on the surface via the MHC-I pathway. Erythrocytic invasion is rapid and propelled by motor proteins [47]. After erythrocytic rupture, the released merozoites recognise new erythrocytes within 1 minute and following recognition and erythrocytic attachment, the parasite completes

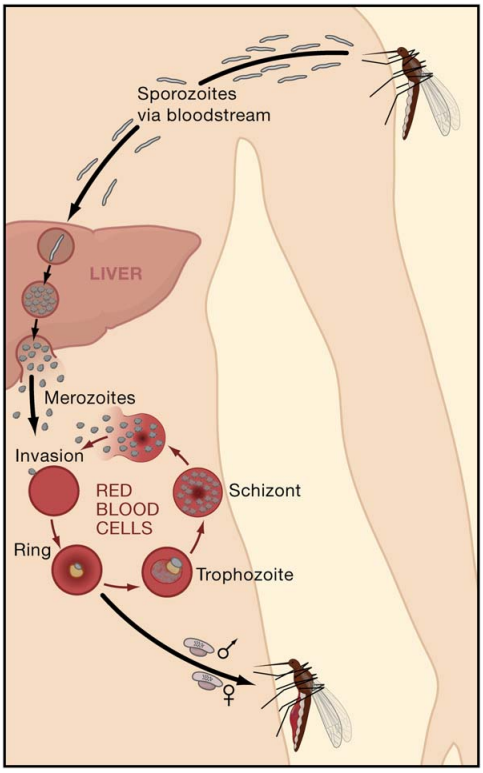


Figure 1.3.2: Lifecycle and pathogenesis of *Plasmodium falciparum* [37]. The female *Anopheles* mosquito is the transmission vector for the malaria parasite. Just after the bite, the mosquito injects saliva containing anti-clotting proteins. Along with the saliva travels the sporozoite form of the malaria parasite. It seeks out the liver, where it will penetrate the cellular wall of liver cells and begin transformation to the next step in the lifecycle. The malaria parasites leave the liver cell as merozoites. The parasite is now ready to begin infection of the red blood cells (RBCs). Every 48 hours a hoard of merozoites are released from each ruptured iRBC, each ready to infect a new RBC. Along with the merozoites, immature male and female gametocytes are released. These will be taken in with the next blood meal of another *Anopheles* mosquito, thus initiating the mosquito life cycle [37].

invasion of the erythrocyte after approximately 27 seconds [59]. The 90 second window of antigenic availability in the merozoite extraerythrocytic stage is therefore highly limited compared to the 48 h intraerythrocytic cycle. The intraerythrocytic stage provide a beneficial niche for merozoite propagation, by providing not only protection from immune detection, but also plenty of nutrition in the form of hemoglobin [32]. As much as 95% of the protein content of an erythrocyte is haemoglobin [52]. The erythrocytic lack of antigen presentation capabilities, may seem odd in this context but several reasons have been proposed as to why erythrocytes lack a nucleus, e.g. oxidative stress avoidance [159], maximise oxygen carrier capacity and cell flexibility for passage of capillary network [79]. To ensure a well functioning population of erythrocytes and thusly an optimised oxygen carrier capability, erythrocytes will, under normal circumstances, undergo senescence after 100-120 days [48]. Erythrocyte membrane changes signify ageing, which macrophages in the spleen will detect and destroy the erythrocyte [56]. This process is known as eryptosis [58]. As with aged erythrocytes, IEs are likewise detected and filtered in the spleen [157]. The induction of eryptosis reduces parasitemia, diseases and genetic traits, resulting in an increase in eryptosis, e.g. sickle cell anemia, therefore provide protection against severe malaria [89]. A crucial step in the pathogenesis of malaria is to avoid this filtering system. The parasite alters the exterior of the IE, expressing surface adhesion proteins [125] known as *Plasmodium falciparum* erythrocytic membrane protein type 1 (PfEMP1) [11]. These proteins facilitate adhesion to a wide variety of host receptors in the microvascular epithelium, e.g. CD36, ICAM-1 and CSA [86]. This way PfEMP-1 mediated cytoadherence facilitates parasitemia, by avoiding splenal destruction of the IE [73].

Unlike the intracellular structures in the IE, the surface adhesion proteins are visible and accessible to the human immune system as small 'knobs' on the surface [26, 146]. The surface displayed antigens, will be recognised by antibodies and high titers of antibodies against IEs are associated with immunity [29]. The antibodies potentially block the PfEMP1:receptor formation thereby

hindering cytoadherence and simultaneously labelling the IE for macrophagocytosis [29]. The blocking is only potential in that non-blocking antibodies are raised against immunodominant B-cell epitopes [72], creating an immunological 'smoke-screen' [90]. The function of this 'smoke-screen' is to divert the immune response away from epitopes located close to the PfEMP1:receptor interaction site. Thereby the IE retains its cytoadherence capability. Furthermore PfEMP1 exhibits an extreme variability [85]. The family of genes encoding the PfEMP1 are known as the var gene family [115]. Each parasite harbours approximately 60 var gene copies [102]. Each parasite express only one PfEMP1 variant at any given time [152]. Parasites expressing a variant, against which no antibodies exists, will be selected for and flourish causing a new wave of parasiteamia [27]. This phenotype switching is known as clonal or antigenic variation [125]. The consequence of the PfEMP1 antigenic variation is that *P falciparum* malaria immunity is gradually acquired after multiple exposures [62]. This gradual acquisition of immunity is primarily antibody based [5], innate immunity however also play a role [144]. The gradual acquisition of immunity follows as the immune system is exposed to different PfEMP1 variants [42]. With each exposure memory b-cells will produced ready to produce anti-adhesion antibodies upon infection [145]. In the context of placental malaria, it is important to note that the acquired immunity does not lead to a sterile immunity, but rather a symptom free infection [142].

Studies have indicated that in some cases two PfEMP1 variants may be simultaneously expressed [21, 80]. In a 2010 study a parasite was found to bind both the CD31/PECAM1 and the CD54/ ICAM1 receptors, the result of which was a 2-fold increase in binding efficiency to human endothelial cells, when compared to parasites expressing only one PfEMP1 variant [80]. This emphasises not only that many aspects of malaria pathogenesis remain non-elucidated, but also the ability of malaria to adapt.

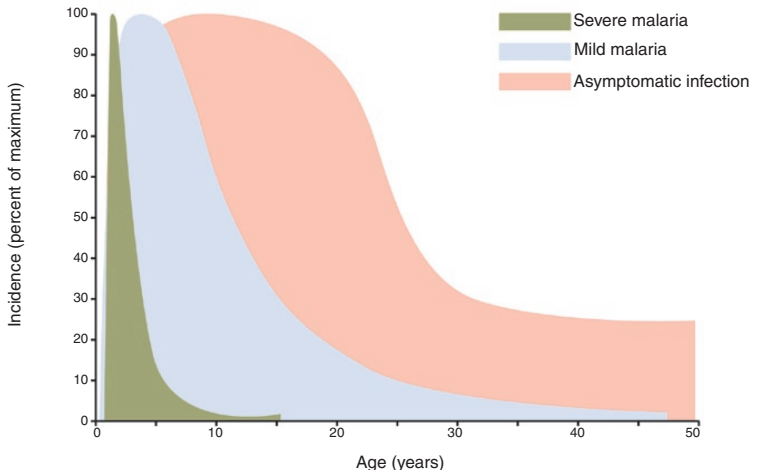


Figure 1.3.3: Gradual acquisiation of immunity to *Plasmodium falciparum* malaria [90]. Malaria immunity is gradually acquired. Infants and children up to the age of 5 experience severe, often fatal, malaria. From the ages of 5 to 10, children experience a milder form of malaria and finally after the age of 10, asymptomatic infections are seen [6, 90, 103].

1.4 Placental malaria

"So long as woman has walked the earth, malaria may have stalked her" [45]. A special case of malaria is pregnancy associated malaria (PAM) or placental malaria. As the name suggests, this type of malaria affects only pregnant women. The clinical consequences of placental malaria include: maternal anaemia, low birth weight, preterm delivery, which leads to increased perinatal and infant mortality [106]. Around 100,000 - 200,000 infants die every year from placental malaria [107]. Women living in malaria-endemic areas will have acquired immunity once they reach child-bearing age, however upon pregnancy a new wave of parasitemia hits [135]. The reason for this sudden set back is the prevalence of novel IE sequestering receptors in the intervillous space of the placenta [76]. The intervillous space is where nutrients and oxygen is exchanged from the maternal blood to the blood of the fetus. Where normal *P. falciparum* primarily utilises the abundantly available human CD36 receptor [73], placental malaria is the result of the interaction of one particular PfEMP1 variant with the

chondroitin sulphate A (CSA) [16] part of placental Chondroitin sulfate proteoglycans (CSPG) [23] found on syncytiotrophoblast cells in the intervillous space of the placenta [1].

The PfEMP1 variant of placental malaria binds exclusively to CSA [24, 75] and since women pre-pregnancy exposure to CSA-binding parasites is limited, the acquired immunity against non-placental malaria, provide no protection against CSA-binding parasites [51]. Thus the infected women fall sick to malaria once again [148]. As with normal malaria, placental malaria immunity is mediated by gradual acquisition of blocking-capable anti-PfEMP1 antibodies and is hence a function of the parity of the woman [42] (See figure 1.4.1). Multigravidae women experience only asymptomatic infection despite high parasitemia [132].

Since placental malaria is a consequence of a very specific receptor:ligand interaction, as described above, the placental PfEMP1 makes for a good vaccine target [17].

1.4.1 PfEMP1 variant VAR2CSA is responsible for placental malaria

In 2004 Salanti and co-workers identified VAR2CSA, a single and uniquely structured molecule belonging to the PfEMP1 family, as the protein responsible for placental malaria [129]. As VAR2CSA is the parasite ligand for placental binding, VAR2CSA is recognised as the leading PAM vaccine candidate. VAR2CSA is a very large 350 kDa [105] multidomain protein [19], divided into 6 duffy binding like domains (DBL) [33]. In total, VAR2CSA consists of approximately 3,000 amino acids residues and equence similarity ranges from 80-98% [39]. VAR2CSA contains both conserved folds and high sequence variation. [69]. VAR2CSA epitopes are predominantly located as conformational epitopes in the polymorphic regions of VAR2CSA [9] and it is likely that the variable regions acts as immunological smoke-screens, directing the immune response towards non-protective epitopes [38, 72, 90]. In 2010 Salanti and co-workers managed to express the entire full-length recombinant VAR2CSA [83]. They also showed that the recombinant VAR2CSA induced potently

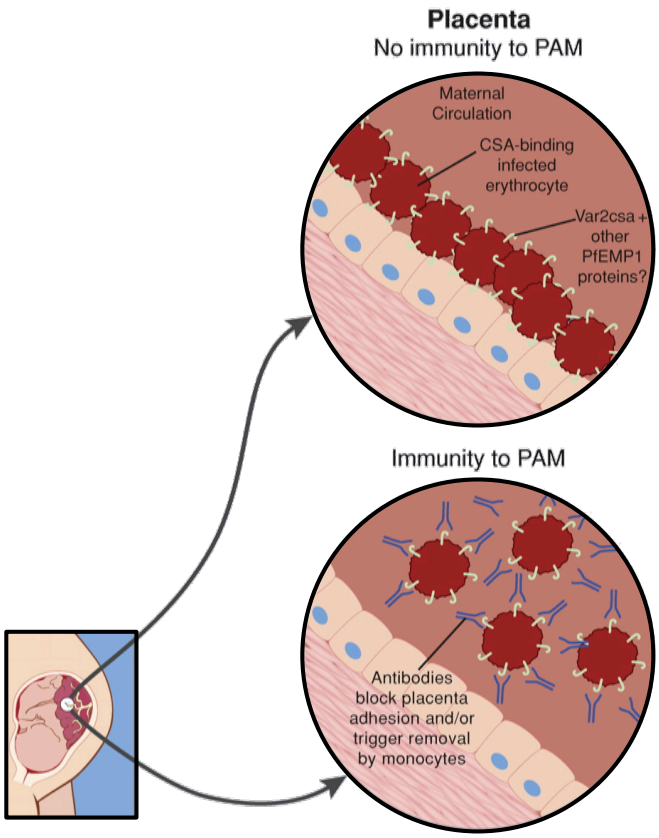


Figure 1.4.1: Placental malaria. Placental malaria arises when otherwise immune women living in areas endemic to malaria, suddenly fall sick to malaria during first pregnancy. The cause of this is the presence of novel receptors, Chondroitin sulfate A (CSA), in the intervillous space of the placenta presented at the syncytiotrophoblast of the endothelial lining. These receptors facilitate sequestration of infected erythrocytes, which causes a new wave of parasitemia, by interacting with the VAR2CSA variant Plasmodium falciparum Erythrocyte Membrane Protein 1 (PfEMP1). Immunity is acquired gradually with increasing parity and is antibody mediated by blocking the CSA:PfEMPP1 interaction. Adapted from [140].

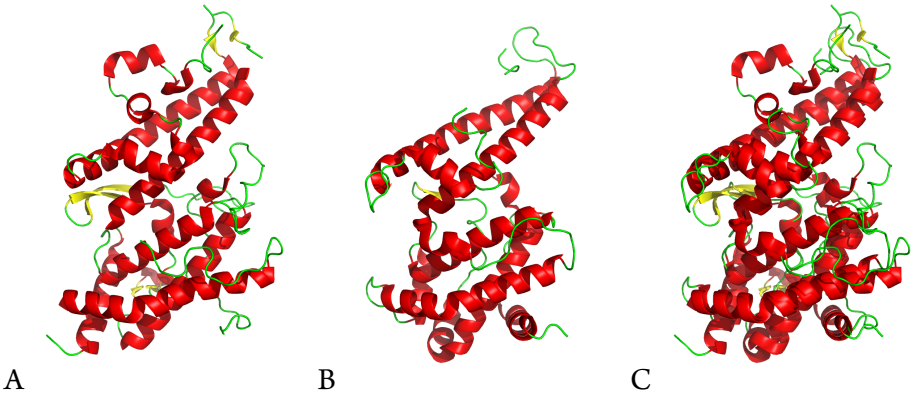


Figure 1.4.2: Crystal structures of VAR2CSA domains **A**: DBL3X at 1.80Å [68] and **B**: DBL6ε at 1.84Å [54] and **C**: DBL3X and DBL6ε aligned. Colour legend: Red = α-helices, green = loops and yellow = β-sheets. The figure illustrates the high similarity of the two VAR2CSA domains.

homologues inhibiting antibodies [83], i.e. when immunised with a given variant, the raised antibodies potently inhibited the same variant. It is however also possible to induce single-domain inhibiting antibodies [7, 110, 130]. Extensive work has gone in to solving the crystal structure of VAR2CSA [33], despite this so far only the single domains DBL3X [68] and DBL6ε [82] have been solved (See figure 1.4.2 for structural comparison).

However the hunt continued by producing truncated versions of VAR2CSA aiming at identifying the smallest possible part of the protein, while still retaining activity, the so called minimum binding region [40]. The minimum binding region is ID1-DBL2Xb (See figure 1.4.3). This region is the minimum part of VAR2CSA required to retain the ability to bind to CSA. As seen in figure 1.4.3 the proposed structure illustrates how the before mentioned immunological smokescreen works by hiding away the active site of the protein and allowing antibodies to adhere to insignificant epitopes located at the exterior of the protein and away from the active site.

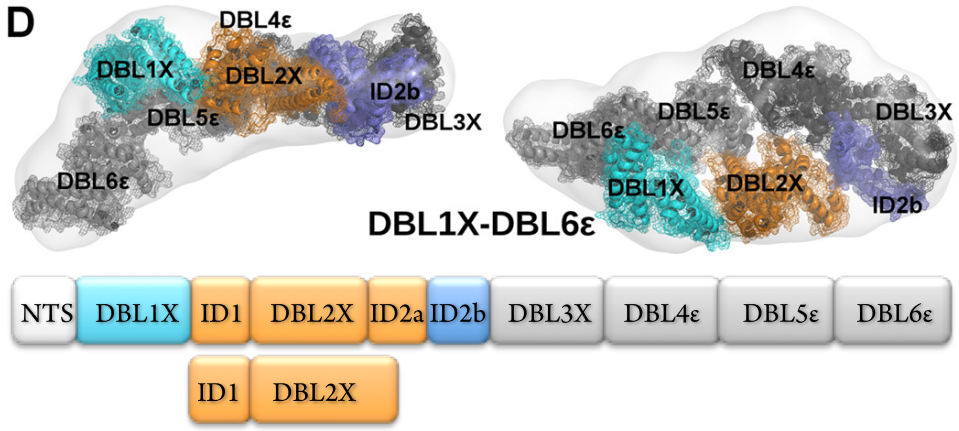


Figure 1.4.3: Proposed structure of VAR2CSA and minimum binding region. Adapted from [33].

Malaria vaccine

The use of experimental human challenges for promoting vaccine development has been proposed [133]. The idea being that the resulting *P. falciparum* infection can be cleared using anti-malarial drugs. The first malaria vaccine, RTS,S/AS01, is expected to be licensed in 2015 [124]. Initial phase III results are promising [2]. The target of the vaccine is the pre-erythrocytic phase of malaria and has been shown to induce potent antibody response against *P. falciparum* circumsporozoite protein (CSP), along with a moderate CD4+ T cell response [49]

Antibody mediated pathogen neutralisation has shown great potential as a therapeutic drug [28]. The downside of using antibodies as a drug, is that the protection only lasts for as long as the antibodies are in the bloodstream. Traditional vaccines on the contrary can potentially induce long lasting antibody mediated immunity [145]. Synthetic approaches, such as RTS,S/AS01, are being applied aiming at mimicking the 'natural' antibody mediated immunity [77]. This protein-plus-adjuvant approach for vaccine development against a specific malaria stage has previously been applied [70].

Henceforth the focus will be on VAR2CSA as vaccine target, it should however be noted that many other targets are available [124]

"Despite the clear importance of PfEMP1 as an immune target, antigenic diversity is a key obstacle that must be overcome for PfEMP1 to be pursued as a major malaria vaccine candidate." [17]

1.5 High Density Peptide Microarrays

1.5.1 Introduction

One approach for the analysis of VAR2CSA sequence variation is by scanning fragments of the VAR2CSA protein (peptides) for biological activity. Using conventional technology peptide scanning is a costly affair, costing around €30 per ten amino acids. The aim of the PepChipOmics projects, was to develop a high throughput chip, lowering the price to a mere few cents per peptide. It may eventually be possible to have 500,000 peptides on a single chip. When dealing with data of such magnitude, the risk of identifying false positives increases dramatically. This calls for the development of robust statistical methods for data analysis. The High Density Peptide Microarray (HDPMa) project, acronym: PepChipOmics, was initiated under the EU's 7th Framework Programme for Research and Technological Development (FP7).

1.5.2 Technical description

The novelty of the HDPMa technology is the use of photo activated amino acid residue coupling. The targeted photo activation is achieved by having a light source shining rays of light onto a collimating lens system. After collimation, all the rays of light are parallel and directed at the Digital Micro mirror Device (DMD) chip. The DMD chip consists of micro mirrors, each capable of rotation and by doing so producing a binary system. The rays are reflected through an imaging lens system onto the peptide chip. This way each field on the peptide chip can be either illuminated or not (See figure 1.5.1).

The DMD chip is no different from the one used in a normal projector used for

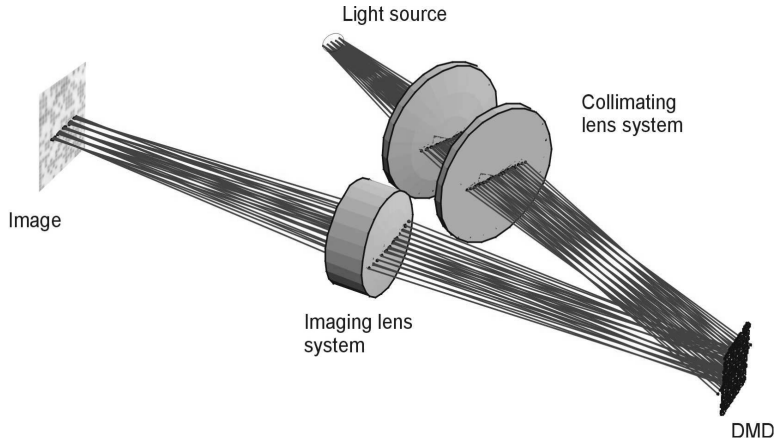


Figure 1.5.1: The DMD chip generates an image, which is projected onto the peptide synthesis support. This determines the location, where the growing peptide is de-protected i.e. prepared for amino acid extension. [Picture from: High-Density Peptide MicroArrays and Parallel On-Line Detection of Peptide-Ligand Interactions (High-Throughput Research in Biotechnology. Acronym: PepChipOmics EU-grant application FP7-HEALTH-2007-B. Coordinator: Søren Buus Professor, MD, PhD. Laboratory of Experimental Immunology, Faculty of Health Sciences, University of Copenhagen, Denmark].

presentations. The DMD chip used is HD, yielding a theoretical capacity of $1920x1080 = 2,073,600$ peptides, however in order to minimise the effect of adjacent coupling fields, blank fields are required in a checkerboard manor. Also the edges are excluded and more than one field with the same peptide is required to estimate signal. Thusly the capacity is reduced to around 200,000 unique peptides.

1.5.3 *In situ* synthesis

The peptides are synthesised *in situ* on the surface of the HDPMa chip. The principle synthesising is similar to that of the standard Fmoc-polyamide synthesis protocol, only the HDPMa chip synthesis utilises photo-sensitive groups for elongation initiation.

1. The chip surface is prepared with c-terminus spacers consisting of 4 aspartic acid residues 'DDDD', lifting the final peptide up from the surface. Each spacer has a photo-sensitive group placed at the N-terminus

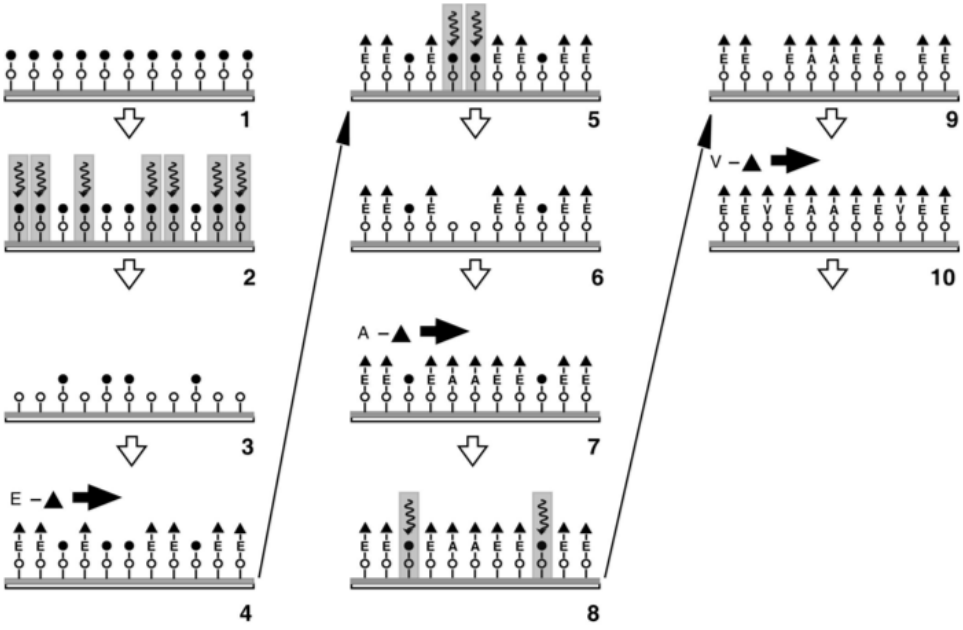


Figure 1.5.2: HDPMa *In situ* peptide synthesis. Filled triangles: Fmoc protection group, filled circles: photosensitive group. Synthesis details in text. [Picture from: High-Density Peptide MicroArrays and Parallel On-Line Detection of Peptide-Ligand Interactions (High-Throughput Research in Biotechnology. Acronym: PepChipOmics EU-grant application FP7-HEALTH-2007-B. Coordinator: Søren Buus Professor, MD, PhD. Laboratory of Experimental Immunology, Faculty of Health Sciences, University of Copenhagen, Denmark].

2. Targeted light rays from the DMD hits target peptides on the HDPMa surface

3. Targeted peptides are de-protected

4. The surface is flushed with an amino acid of choice, coupling the amino acid to the de-protected N-terminus of the target peptide

5. Steps 5-10 are repeats of steps 2-4. Lastly the Fmoc groups (filled triangles) are exchanged with photosensitive groups (filled circles)

1.5.4 Epitope mapping using HDPMa

One application of the HDPMa technology is the scan for linear B-cell epitopes in proteins (See section 1.2 for epitope explanation). A given protein can be traversed using a window of 15 amino acid residues (creating 15-mers). Each of the 15-mers is then synthesised *in situ* onto a spacer on the peptide support. The spacer is used to increase the availability of each peptide. For the VAR2CSA scan a spacer of 4 aspartic acid residues was used. Once the chip is fully synthesised, it is incubated with relevant fluorophor conjugated primary antibodies. (see fig1.5.3A) and the chip, with potentially fluorescent fields (see fig. 1.5.3B), is photographed.

The HDPMa chip was in this case constructed in a manor similar to that of a chequerboard, where blank fields surrounds the actual peptide field. This allows the signal from each peptide field to be normalised with that of the blank fields surrounding it. Thus resulting in 15-mer sequences and associated signal-to-noise ratio (S/N). The signal is quantified using the red-green-blue (RGB) colour model, in which (0,0,0) is black and (255,255,255) is white. This way 256 shades from black to white can be quantified.

1.5.5 HDPMa development

The technology is not fully developed and is therefore subject to further development. The latest addition is the use of 2x2 mirror setup per field [25]. This reduces the risk of having a dead field. Also field spaces have been eliminated increasing the capacity to $\sim 518,000$ peptides. For this study the system described above was utilised. However since this study, the antibody detection system has been altered, such that the peptide:antibody binding is quantified using a secondary fluorophor conjugated antibody.

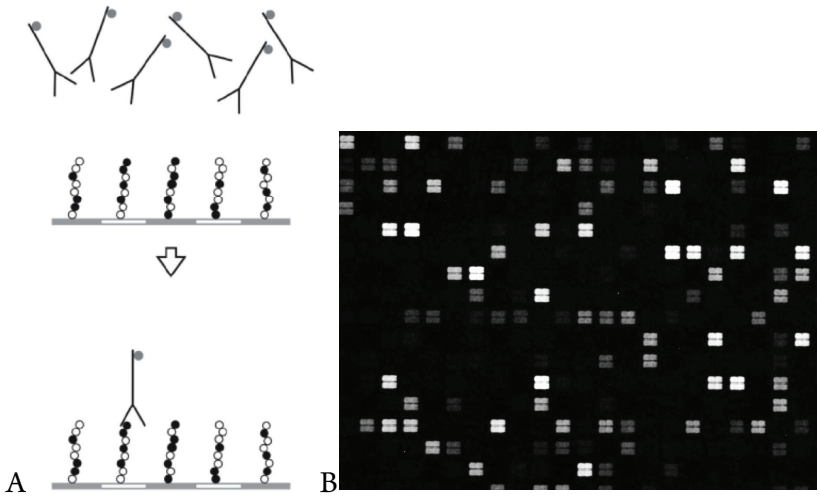


Figure 1.5.3: A: Peptide scan of linear B-cell epitopes by method of fluorophor conjugated antibody interaction. The fully synthesised peptide chip is incubated with fluorophor conjugated antibodies raised against the native protein. After washing, the fluorescence is recorded. **B**: Example of peptide chip image[64]. Each 2x2 square represents the signal from a peptide single field. The raw signal can then be read by computing the mean RGB value of the field ($s_f \in [0; 255]$). [Picture from: High-Density Peptide MicroArrays and Parallel On-Line Detection of Peptide-Ligand Interactions (High-Throughput Research in Biotechnology. Acronym: PepChipOmics EU-grant application FP7-HEALTH-2007-B. Coordinator: Søren Buus Professor, MD, PhD. Laboratory of Experimental Immunology, Faculty of Health Sciences, University of Copenhagen, Denmark].

1.6 On statistics

Introduction and background on the statistics used in this thesis is given in context with part II.

Part II

Paper I: SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments

There are three kinds of lies: lies, damned lies, and statistics.

Mark Twain

2 SigniSite

2.1 Brief introduction to *SigniSite*

The challenge of genotype-phenotype correlation can be stated as follow: *"Given a protein of interest, with a phenotype, which can be numerically quantified, e.g. the binding affinity, catalytic efficiency, fluorescent signal, immunogenicity or similar, which residue(s) constitute the genotype determining the observable phenotype?"*. *SigniSite* addresses this challenge by analysing a multiple sequence alignment (MSA) of variants of a protein of interest, where each variant is associated with numerical value quantifying a phenotype of interest. The phenotype determining residues are then identified by means of a non-parametric mean rank-test analysis of each unique residue at each position in the MSA.

Within this framework each unique residue at each position constitute a test. In order to address the multiple testing scenario originating from this framework,

either the 'Holm step-down' or 'Bonferroni single-step' [44] methods for correction for multiple testing (CMT) can be applied.

However as *SigniSite* was intended to (also) be used by experimentalists performing wet-lab mutation-analysis, it is envisioned that relatively small MSAs could be submitted to *SigniSite*. One down-side of using CMT is that while the risk of false positives is markedly reduced, the risk of false negatives is markedly increased. In other words applying a CMT, which is too harsh, may in the end quench the significance of more subtle genotype-phenotype correlations. The challenge of controlling the 'harshness' of the correction, is the offset for the work described in this part of the thesis. Moreover:

"Is it possible to reduce the total number of tests by identifying and eliminating any superfluous tests based on an unbiased evaluation of the system prior to SigniSite analysis?"

2.2 Briefly on the *SigniSite* method

Details can be found in the *SigniSite* paper and supplementary materials. As prerequisite for the *SigniSite* analysis, each sequences in the submitted MSA must be associated with a value $v \in \mathbb{R}$ representing the phenotype to be analysed. *SigniSite* sorts and ranks the sequences based on v. Then for each unique amino acid residue *aa* present at each position p_i in the MSA, *SigniSite* performs a non-parametric test by computing the observed mean rank, $\overline{x}_{obs}$, and comparing it to the expected mean rank μ_{exp}, given that we expect that the residue aa_{p_i} has no preference for neither high nor low phenotypic values and therefore is expected to be uniformly distributed at a given position p_i, $i = \{1, 2, \ldots, m\}$, in a MSA of length m. The expected mean is given by:

$$\mu_{exp} = \frac{N+1}{2} \tag{2.1}$$

Where N is the number of sequences in the MSA. The standard deviation can be approximated by:

$$\sigma = \sqrt{\frac{(N - n_{aa})(N + 1)}{12 n_{aa}}} \tag{2.2}$$

Where n_{aa} is the number of unique amino acid residues aa, observed at p_i. The parameters μ_{exp} and σ describes the expected distribution. We can therefore calculate a *z-score* representing the probability of observing $\bar{x}_{obs}$ or a value 'more extreme', given that μ_{exp} is the 'true' mean:

$$z = \frac{\mu_{exp} - \bar{x}_{obs}}{\sigma} \tag{2.3}$$

This obtained *z-scores* can be approximated by the normal distribution and *p-values* can therefore be obtained by standard method.

2.3 Introduction to multiple testing

For any position p_i in a given MSA, where at least two amino acid residues are present, *SigniSite* will perform one test for each unique residue. From this it is obvious, that multiple tests are performed. However, as multiple tests are performed, the likelihood of identifying 'something' as being significant by chance, rather than 'actual' significance increases. This can be addressed by applying what is known as 'Correction for Multiple Testing' (CMT).

In classical statistics the first step is to state a null-hypothesis and an alternative hypothesis. Within the *SigniSite* framework, the null-hypothesis H_0 is:

$$H_0 : \bar{x}_{obs} = \mu_{exp} \tag{2.4}$$

and the alternative hypothesis H_1 is:

$$H_1 : \bar{x}_{obs} \neq \mu_{exp} \tag{2.5}$$

Subsequently either the null-hypothesis or the alternative hypothesis is rejected, based on a chosen level of significance $0 < \alpha < 1$. By convention usually 5% or 1%, referred to a *'significant'* or *'highly significant'* respectively. The level of significance is used as a 'critical value' to evaluate the outcome of a test quantifying the probability of observing a value as, or more extreme than the tested value, given that the null-hypothesis is true. The result of this tests is known as the *p-value*. If the *p-value* is less than or equal to the level of significance α, i.e. $p \leq \alpha$, then the null-hypothesis is rejected. If $p > \alpha$, then the null-hypothesis cannot be rejected. Note, neither the null- or alternative hypothesis can be accepted, only rejected or not rejected. Given the *SigniSite* framework, we do not know beforehand whether $\overline{x}_{obs} < \mu_{exp}$ or $\overline{x}_{obs} > \mu_{exp}$, we therefore clearly need a two-sided hypothesis test. In a two-sided hypothesis test it is evaluated if $p < \frac{\alpha}{2}$. In either case, when conducting such tests, two types of errors can occur:

1. Type I error occurs, when a true null-hypothesis is incorrectly rejected. This type of error is also know as a 'false positive'.

2. Type II error occurs, when a one fails to reject a false null-hypothesis. This type of error is also know as a 'false negative'.

When performing multiple tests, the risk of type I errors increase, CMT is introduced in order decrease this risk. Often the default way of performing CMT is using the *'Bonferroni single-step'* method for CMT, in which the adjusted *p-value*, p_{adj} is obtained by multiplying the *p-value* with the number of tests performed n_t:

$$p_{adj} = min\left(1, p \cdot n_t\right) \tag{2.6}$$

Another method often used is the *'Holm step-down'* method for CMT, in which a set of k obtained *p-values* $\{p_1, p_2, \ldots, p_k\}$ are sorted ascending and the lowest, i.e. the strongest, *p-value* is multiplied with n_t, the second lowest with $n_t - 1$ and so on, yielding a vector $\mathbf{P}_{adj}$:

$$\mathbf{P}_{adj} = \left(min\left(1, p_i \cdot n_t\right), min\left(1, p_{i+1} \cdot \left(n_t - 1\right)\right), \ldots, min\left(1, p_m \cdot 1\right)\right) \tag{2.7}$$

A given *p-value* is subsequently considered significant, if p_{adj} is smaller than the chosen level of significance α, i.e.

$$p_{adj} \leq \alpha \tag{2.8}$$

It should be noted that in both the 'Bonferroni single-step' and the 'Holm step-down' method for CMT, the most significant/strongest *p-value* is multiplied with the total number of tests n_t. Therefore, when looking at the lowest *p-value* in the case of the 'Holm step-down' method and all *p-values* in the case of 'Bonferroni single-step' method, we can adjust the α-value with the number of tests n_t rather than the *p-value* and eq. 2.8 becomes:

$$p_{adj} \leq \alpha \Rightarrow min\left(1, p \cdot n_t\right) \leq \alpha \Rightarrow p \leq \frac{\alpha}{n_t} = \alpha_{adj} \tag{2.9}$$

The min() function is lifted, as $0 < \alpha < 1, n_t > 0 \Rightarrow \alpha_{adj} < 1$. Using eq. 2.9, we can take a closer look at how CMT impacts the significance threshold for the strongest obtained *p-value* in a multiple testing scenario.

2.4 CMT IMPACT ON SIGNIFICANCE THRESHOLD

We can evaluate the impact of the Bonferroni and Holm methods on the strongest *p-value* by looking at both the adjusted α value, α_{adj}, and the corresponding adjusted *z-score* threshold, $z_{adj} = z_{1-\alpha/n_t}$ (see eq. 2.9). Here the function $z(\alpha_{adj})$ refers to the normal distribution $\mathcal{N}(\mu = 0, \sigma^2 = 1)$. In order to quantify the CMT impact, we define a vector $\mathbf{n}$, from which $\alpha_{adj}(\mathbf{n})$ and subsequently $z(\alpha_{adj}(\mathbf{n}))$ follow:

$$\mathbf{n} = [1, 2, .., 100] \Rightarrow \tag{2.10}$$

$$\alpha_{adj}(\mathbf{n}) = [0.05, 0.025, ..., 0.0005] \Rightarrow \tag{2.11}$$

$$z(\alpha_{adj}(\mathbf{n})) = [1.96, 2.24, ..., 3.48] \tag{2.12}$$

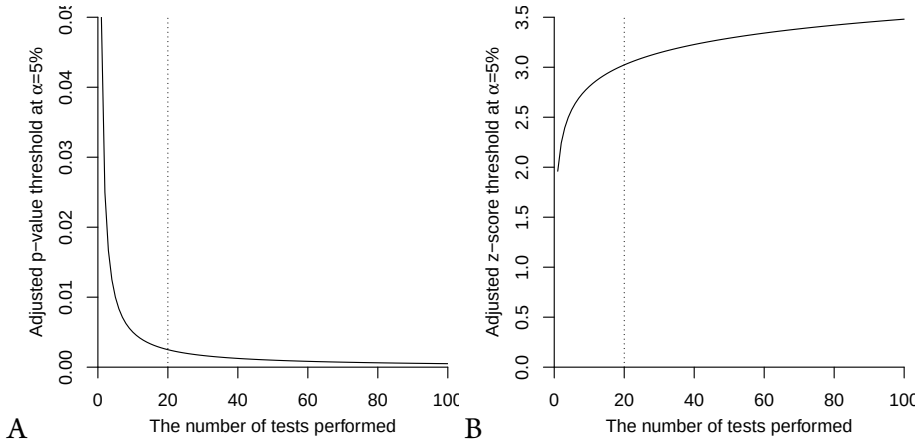


Figure 2.4.1: Bonferroni corrected $|z|$ and p thresholds as a function of the number of tests performed n_t. **A**: $p_{adj}(n_t)$. **B**: $|z_{adj}(n_t)|$.

We can now depict the functions $a_{adj}(\mathbf{n})$ and $z(a_{adj}(\mathbf{n}))$ (See fig. 2.4.1). From this, it is evident that CMT strongly impact the signficance threshold and especially with the first approximately 20 tests, as fig. 2.4.1 clearly shows. Calculating that:

$$\begin{aligned} a_{adj}(1, 20, 100) &= (0.05, 0.0025, 0.0005) \\ z(a_{adj}(1, 20, 100)) &= (1.96, 3.02, 3.48) \end{aligned}$$

We can calculate that the first 20% of the tests constitute 70% of the total increase in the *z-score* threshold, z_{adj} and 96% of the total decrease in the *p-value* threshold a_{adj}.

SigniSite performs molecular-scale genotype-phenotype inference. In smaller alignments, the conservative CMT, may quench the significance of a subtle genotype-phenotype association. In more general terms, CMT increases the risk of type II errors. However as we will discuss in the following the number of tests, which controls the 'harshness' of the CMT, can be reduced.

2.5 CMT of inhomogeneous systems

The purpose of applying CMT, is to reduce the risk of type I errors, by increasing the significance threshold. However a consequence of increasing the significance threshold is that the risk of type II errors is increased. The threshold adjustment is a trade-off between type I and type II errors.

Both the 'Bonferroni single-step' and the 'Holm step-down' method for CMT, assumes that the system being tested is homogenous with respect to the information content of each test. However this is far from always the case. The problem can be illustrated using the following 'flip-a-coin' example: We wish to test 40 coins aiming at identifying any unfair coins. Our null-hypothesis is that the coin is fair, i.e. 50/50 chance of heads/tail. The first coin is flipped 20 times, producing 16 heads, under the null-hypothesis the probability of this is $p = 0.0046$, the second is flipped 20 times, producing 15 heads, corresponding to $p = 0.015$. However the remaining 38 coins are only flipped twice. Nevertheless the number of tests is 40, yielding CMT *p-values* of $p = 0.18$ and $p = 0.59$ respectively. Had we been able to identify the 38 tests with only 2 coin flips as superfluous, the CMT *p-values* would instead have been $p = 0.0092$ and $p = 0.030$ respectively.

The 'flip-a-coin' system is inhomogeneous in that we have much more information from the first two coin flips than the following 38. Adjusting for multiple testing, the first coins cannot be significantly identified. When considering the uneven distribution of information in the system, it intuitively seems too harsh to adjust the trial of the first coins this way. If we can somehow identify which test should be performed and which should not before performing the actual testing, we would, in this case, be able to identify the first two coins as being unfair, which seems intuitively likely.

The inhomogeneous system content in the 'flip-a-coin' example is analogous to a MSA. In a MSA the amino acid distributions will differ between positions, i.e. both the number of unique residues at a position and the frequency of these will differ. From this it is therefore obvious that a MSA constitute an inhomogeneous

system and that some positions will have a higher information content than others.

2.6 Estimating MSA positional information content

One approach for reducing the number of tests, could be filtering based on the positional information content prior to performing the statistical evaluation. Thus as previously mentioned, aiming to reduce the number of tests performed and thereby allowing identification of more subtle associations. The assumption was that highly conserved MSA positions were less likely to be impacting the observable phenotype.

The proposed approach was to calculate the positional information content and then choosing a conservation cut-off for whether or not to include the MSA position in the set of test. This is the approach, which was utilised in the analysis of the VAR2CSA-dbl5ε sequence analysis [61].

Traditionally the positional information content, for a given position p_i in a MSA is evaluated using the following 3 measures [99]: *i.* the Shannon entropy [137] $\mathrm{H}(p_i)$, *ii.* the Kullback-Leibler divergence [88] $D(p_i||q)$ and *iii.* the Shannon information content I:

$$\mathrm{H}(p_i) = -\sum_a p_a log_2(p_a) \tag{2.13}$$

$$D(p_i||q) = \sum_a p_a log_2\left(\frac{p_a}{q_a}\right) \tag{2.14}$$

$$I(p_i) = log_2(20) + \sum_a p_a log_2(p_a) \tag{2.15}$$

Where a denotes all unique amino acid residues present at p_i, p_a is the frequency of residue type a at p_i and q_a is the global background frequency for residue type a. If a given p_i is completely conserved, i.e. only one type of residue is present, then $\mathrm{H}(p_i) = 0$. By convention $0 < \mathrm{H}(p_i) \leq 1$ is considered 'highly conserved' [57], $1 < \mathrm{H}(p_i) \leq 2$ is conserved and $2 < \mathrm{H}(p_i) \leq 4.3$ is variable. It should be

noted that since $I(p_i) = log_2(20) - \mathrm{H}(p_i)$, the inversed cut-offs are applicable for $I(p_i)$. However choosing to omit position based on a 'highly conserved' cut-off posed a problem in that the maximum obtainable *SigniSite z-score* for an amino acid residue *aa* at position p_i in a MSA of N sequences occurs, when $n_{aa} = \frac{N}{2}$ (will be shown later). E.g. given a position p_i, with 2 unique residues, for which the frequency $f_{aa_1} = f_{aa_2} = 0.5$ then the Shannon entropy would be:

$$\mathrm{H}(p) = -\sum_a p_a log_2(p_a) = -2\left(\frac{1}{2} log2\left(\frac{1}{2}\right)\right) = 1 \tag{2.16}$$

If we then chose to omit 'highly conserved positions', i.e. by convention, positions for which $0 < \mathrm{H}(p_i) \leq 1$, we would omit position harbouring the potential of obtaining the maximum possible *SigniSite z-score*. It was therefore concluded that given the *SigniSite* framework, neither $\mathrm{H}(p_i)$ nor $I(p)$ provided a good measure for pre-test filtering based on the positional degree of conservation. With respect to $D(p||q)$, then $D(p||q)$ is equal to $I(p)$, when $q = \frac{1}{20}$ for all types of residues. The approach of positional exclusion based on conventional evaluation of information content was therefore abandoned and we instead turned to investigate if we could find an intrinsic property of the *SigniSite* framework, which could provide us with a measure for determining whether or not to include a given residue in the set of tests to be performed.

2.7 Analysis of *SigniSite* framework

First let us start with evaluating the maximum obtainable *z-score*. The maximum obtainable *z-score* must occur, when the distance δ between the expected mean rank μ_{exp} and the observed mean rank $\overline{x}_{obs}$ reaches its maximum δ_{max}: For a given number n_{aa} of a given amino acid residue *aa* at p_i, δ_{max} occurs if the occupied ranks of *aa* are $r_{aa,p_i} = \{1, 2, \ldots, n_{aa}\}$, i.e. all the residues of type *a* are top-ranked (or bottom-ranked), fig. 2.7.1 illustrates this situation.

The maximum obtainable mean rank can then be calculated similarly to the

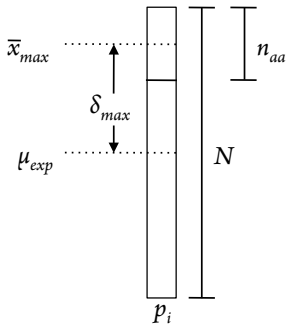


Figure 2.7.1: Graphical representation of δ_{max}. $\overline{x}_{max}$ is the maximum observable mean rank, μ_{exp} is the expected mean rank, N is the number of sequences in the MSA, n_{aa} is the number of a given unique amino acid residue at a given position p_i in a MSA. $\overline{x}_{max}$ occurs when all the residues of a given type are 'top-ranked' (or bottom ranked).

calculation of μ_{exp}, i.e.:

$$\overline{x}_{max}(n_{aa}) = \frac{n_{aa}+1}{2} \tag{2.17}$$

Since the *SigniSite z-score* quantifies the distance from the expected rank, μ_{exp}, to the observed rank, $\overline{x}_{obs}$, then the maximum obtainable *z-score* must occur, when the distance between μ_{exp} and $\overline{x}_{obs}$ is as large as possible and this is the case, when:

$$\overline{x}_{obs} = \overline{x}_{max} = \frac{n_{aa}+1}{2} \tag{2.18}$$

Recalling that $z = \frac{\mu_{exp}-\overline{x}_{obs}}{\sigma}$, we can derive an expression for $z_{max}(n_{aa}, N)$, by combining eqn. 2.17 and eqn. 2.3:

$$z_{max} = \frac{\mu - \overline{x}_{max}}{\sigma} = \frac{\frac{N+1}{2} - \frac{n_{aa}+1}{2}}{\sqrt{\frac{(N-n_{aa})(N+1)}{12n_{aa}}}} = \sqrt{\frac{3n_{aa}(N-n_{aa})}{N+1}} \tag{2.19}$$

If we then choose $N = 100$, we can plot $z_{max}(n_{aa}, N = 100)$ recalling that $0 < n_{aa} < N$ we get the depiction in fig. 2.7.2. Looking at fig. 2.7.2 it becomes clear that the maximum of the computed $z_{max}(n_{aa})$ values occurs, when $n_{aa} = \frac{N}{2}$. This can be shown, by first differentiating eqn. 2.19 with respect to n_{aa} at

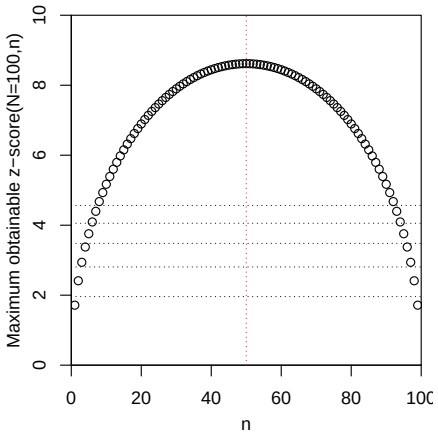


Figure 2.7.2: $z_{max}(n_{aa})$ for $N = 100$. On the x-axis is $0 < n_{aa} < N$ and on the y-axis it $z_{max}(n_{aa})$. The vertical red dotted line marks $max\left(z_{max}(n_{aa})\right)$. The horizontal black dotted lines marks the adjusted *z-score* threshold z_{adj} for $n_t = (1, 10, 100, 1000, 10000) \Rightarrow z_{adj} = (1.96, 2.81, 3.48, 4.06, 4.56)$.

constant N:

$$z'_{max} = \left(\sqrt{\frac{3n_{aa}(N - n_{aa})}{N+1}}\right)' = \frac{3(N - 2n_{aa})}{\sqrt{12n_{aa}(N - n_{aa})(N+1)}} \tag{2.20}$$

Then the function optimum is found by solving the equation for n_{aa}, when $z'_{max} = 0$:

$$z'_{max} = 0 \Rightarrow \frac{3(N - 2n_{aa})}{\sqrt{12n_{aa}(N - n_{aa})(N+1)}} = 0 \Leftrightarrow n = \frac{N}{2} \tag{2.21}$$

So the optimal conditions for a given set of residues at a given position are when all residues are top-ranked and n_{aa} is $\frac{N}{2}$.

Furthermore the depiction in fig. 2.7.2 show how whether or not a given residue can obtain significance under optimal conditions is a function of that particular residue and the number of tests performed, e.g. in a MSA where $N = 100$ and $n_{aa} = 1$ or $n_{aa} = 99$ (fig. 2.7.2 clearly show the symmetry around $\frac{N}{2}$), the maximum obtainable *a-score* is $z_{max} = 1.71$ (See eqn. 2.19). This means that if the count of one unique residue is $n_{aa} = 1$ in a MSA of 100 sequences and

that the level of significance of $95\% \Rightarrow a = 0.05 \Rightarrow z = 1.96$, then this residue can never obtain a significant *p-value* even if we do not perform CMT (See fig. 2.7.2, the lower horizontal dotted line). If we perform CMT at a $10,000$ tests, then $z = 1.96 \Rightarrow z_{adj} = 4.56$ and the prerequisite for n_{aa} in order to obtain significance under optimal conditions become $8 < n_{aa} < 92$ (See fig. 2.7.2, the upper horizontal dotted line). This means that performing a test in either of these cases for the particular residue regardless of the observed rank is completely superfluous. The observation, that a given residue never can obtain significance even under optimal conditions spawned the idea that a residue count-threshold based on the number of sequences and the number of tests could be used to reduce the total number of tests being performed, by filtering these low-count residues. The challenge was to identify the magnitude of the threshold.

2.8 Filtering on n_{aa} as a function of N and n_t

The idea of a residue count-threshold can be illustrated by choosing a threshold t, as illustrated in eqn. 2.22:

$$0 + t < n_{aa} < N - t \tag{2.22}$$

The challenge however is how to estimate the value of the parameter t. In any given alignment with m positions, we know the number of sequences N and the total number of tests can be calculated (regardless of associated values) as the sum of the number of unique amino acid residues n_{aa}^{unique} at each position p_i, skipping completely conserved positions, i.e. when $n_{aa} = N$:

$$\sum_{i=1}^{m} n_{aa,p_i}^{unique} \mid \left(n_{aa,p_i}^{unique} \neq N\right) \tag{2.23}$$

Knowing the number of tests, we can calculate the adjusted *z-score* threshold. Given the corrected threshold, we can calculate the minimum (and maximum) number of residues required to obtain a *z-score* larger than or equal to the

adjusted threshold. This is done by solving letting $z_{max} = z_{adj}$ and $n = n_{thres}$ and then solving for n_{thres} eq. 2.19 for n:

$$\begin{aligned} z_{max} &= \sqrt{\frac{3n(N-n)}{N+1}} \Rightarrow \\ z_{adj} &= \sqrt{\frac{3n_{thres}(N-n_{thres})}{N+1}} \Rightarrow \\ 0 &= -3n_{thres}^2 + 3Nn_{thres} - z_{adj}^2(N+1) \end{aligned}$$

Solving this by aid of a computer algebra system (CAS) tool yields:

$$n_{thres}(N, z_{adj})_{lower} = \frac{1}{2}N - \frac{1}{6}\sqrt{(-12Nz_{adj}^2 + 9N^2 - 12z_{adj}^2)} \quad (2.24)$$

$$n_{thres}(N, z_{adj})_{upper} = \frac{1}{2}N + \frac{1}{6}\sqrt{(-12Nz_{adj}^2 + 9N^2 - 12z_{adj}^2)} \quad (2.25)$$

Where 'lower' and 'upper' denote the minimum and maximum boundaries for whether or not a given residue should be included. It should be noted that as expected $n_{thres}(N, z_{adj})_{upper} = N - n_{thres}(N, z_{adj})_{lower}$, which corresponds with the symmetry we observed in fig. 2.7.2. This way we can make an evaluation of the minimum and maximum number a given residue must have to be included in the final statistical analysis, based on the number of sequences N in the MSA and the number of tests n_t, by combining eqn. 2.22 with the solution:

$$t < n_{aa} < N - t$$

Where:

$$t = \frac{1}{2}N - \frac{1}{6}\sqrt{(-12Nz_{adj}^2 + 9N^2 - 12z_{adj}^2)} \quad (2.26)$$

Figure 2.8.1 illustrates eqn. 2.8, for an alignment of $N = 20$ sequences and where the number of tests is $n_t = 20$ for fig. 2.8.1A and $n_t = 300$ for fig. 2.8.1B:

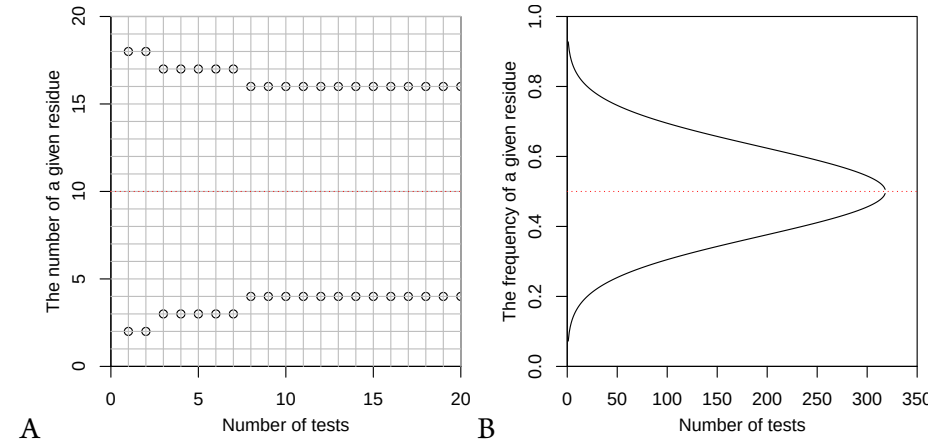


Figure 2.8.1: Depiction of the lower and upper count boundaries for inclusion of a given amino acid residue aa in the set of tests, as given by eqn. 2.8, when the number of sequences in the MSA is $N = 20$. **A**: On the x-axis is $n_t = (1, 2, ..., 20)$. Each number of tests corresponds to an adjusted *z-score* threshold, $z_{adj}(n_t) = (1.96, 2.24, ..., 3.02)$, where the initial level of significance is 95%, whereby $\alpha = 0.05$ and $z_{1-\alpha/2} = 1.96$. On the y-axis is the lower and upper boundaries of $n_{aa} \in \mathbb{N}$ for inclusion in set of tests, e.g. if 10 tests are performed any residues aa at p_i for which $3 < n_{aa} < 17$ will be included in the set of tests to be performed. The red line represent the optimal n_{aa}, i.e. $n_{aa} = \frac{N}{2}$. **B**: Like A, except $n_t = (1, 2, ..., 350)$ and the values on y-axis are given as frequencies $f \in \mathbb{R}$. Here the red line marks $f_{opt} = 0.5$.

2.9 Filtering on n_{aa} as a function of N

When conducting clinical trials, where e.g. the effect of a particular drug is to be investigated, a common challenge is how many subjects to enrol in the trial. If the sample size is too small, it may not be possible to identify a given drug effect as significant, even though there truly is an effect. It is therefore important to estimate the necessary sample size in order to identify a 'meaningful difference' between two groups. One might be in a situation, where it will be simply be too costly to involve the required number of people. Of course it is more convenient to come to this conclusion, before beginning the trial.

In this context, it is important to note that when computing *p-values* by comparing two samples assumed normally distributed, *any* (!) difference, no matter how small, can be identified as significant, given enough data - Hence the before mentioned 'meaningful difference'.

Given two samples, where we assume that there is truly a difference between the two samples, then the power of the test relates to the likelihood of identifying that difference. If the sample size is too small, then the test cannot identify the difference no matter how significant it may be. From this we can estimate how large a sample is needed if we wish to be able to identify a 'meaningful difference' between two samples of interest. This can be extrapolated to the analysis performed by *SigniSite*, in that given a position, where 'too few' resides are present, we can estimate if the sample size is large enough for us to find a significant difference. If this is not the case, then the test is superfluous. In order to estimate the necessary sample size, we introduce the concept of 'power of a statistical test'. Recall that within the *SigniSite* framework, the null-hypothesis H_0 is:

$$H_0 : \overline{x}_{obs} = \mu_{exp} \tag{2.27}$$

and the alternative hypothesis H_1 is:

$$H_1 : \overline{x}_{obs} \neq \mu_{exp} \tag{2.28}$$

We do not know beforehand whether $\bar{x}_{obs} < \mu_{exp}$ or $\bar{x}_{obs} > \mu_{exp}$, therefore we clearly need a two-sided hypothesis test. The test will be conducted at a level of significance of α. This means that H_0 cannot be rejected if:

$$\mu + \sigma \cdot z_{\alpha/2} \leq \bar{x}_{obs} \leq \mu + \sigma \cdot z_{1-\alpha/2} \tag{2.29}$$

Note that if $\alpha = 0.05$, then $z_{\alpha/2} = -1.96$ and $z_{1-\alpha/2} = 1.96$. For any other value of $\bar{x}_{obs}$, H_0 will be rejected. Let us say, that the alternative hypothesis H_1 is in fact true, then it would be convenient to have some idea as to what we would like the probability of rejecting H_0 to be. This probability of rejecting H_0, given that H_1 is actually true is the power of the test. The power of the test is denoted $1 - \beta$ and typically a power of 80% is chosen, for which $\beta = 0.20$. For the sake of maintaining overview of α and β values in relation to hypothesis testing, table 2.9.1 summarises, what has been described so far.

Actual	Pred.	Category	Error type	Probability	Value
1	1	True positive	-	$P(R(H_0)\|H_0 = F)$	$= 1 - \beta$
1	0	False negative	Type II	$P(A(H_0)\|H_0 = F)$	$= \beta$
0	1	False Positive	Type I	$P(R(H_0)\|H_0 = T)$	$= \alpha$
0	0	True Negative	-	$P(A(H_0)\|H_0 = T)$	$= 1 - \alpha$

Table 2.9.1: Overview of α and β values in relation to hypothesis testing. α is the probability of a type I error and referred to as the level of significance. β is the probability of a type II error. The power of a test is defined as $1 - \beta$. Common values are $\alpha = 0.05$ and $\beta = 0.20$. The probability syntax is $P(decision|actual)$, where R is 'reject', A is 'not-reject', T is 'true' and F is 'false'. Adapted from [127].

Choosing sample size relates to the power of test, in that we wish to estimate the sample size needed to be able to detect a significant difference with a probability of $1 - \beta$. Fig. 2.9.1 illustrates the situation for $\bar{x}_{obs} = \bar{x}_{max} < \mu_{exp}$. Here, H_0 cannot be rejected if:

$$\mu + \frac{\sigma \cdot z_{\alpha/2}}{\sqrt{n}} \leq \bar{x}_{obs} \leq \frac{\sigma \cdot z_{1-\alpha/2}}{\sqrt{n}} \tag{2.30}$$

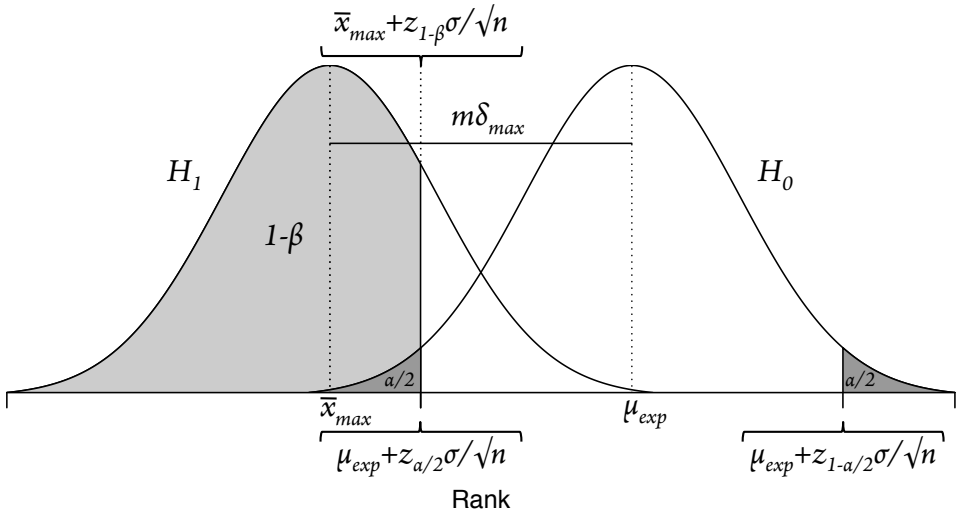


Figure 2.9.1: Plot of distributions of expected and observed rank. $\overline{x}_{max}$ is the maximum obtainable of the observed mean rank, $m \in]0, 1]$ determines the fraction of δ_{max}, which is not to be overlooked, i.e. the 'meaningful difference' to be identified, μ_{exp} is the expected mean rank, $\delta_{max} = |\mu_{exp} - \overline{x}_{max}|$ is the maximum possible difference between the observed and expected mean rank, $z_{a/2}$ and $z_{1-a/2}$ are the critical values for a two-sided level of significance, the dark grey shaded areas are $a/2$, $z_{1-\beta}$ is the critical value for the power of the test and the light grey shaded area is $1 - \beta$. Adapted from [127]

And as before, H_0 will be rejected for any other value of $\overline{x}_{obs}$. Note that the standard error *SE* is $\sigma/\sqrt{n}$. The *SE* is used here since we are sampling multiple times, each time calculating the mean, the standard deviation *SD* of these calculated means is the *SE*. The 'location' of this critical value determining whether to reject or not reject H_0, is controlled by the level of significance a, the power of the test however relates to the area under the distribution of the alternative hypothesis, H_1, to the left of $\mu_{exp} + z_{a/2}\sigma/\sqrt{n}$ and it is this area we want to control by choosing a certain power, e.g. 80% or 90%. So the challenge is, choosing a value of $m \in]0; 1]$, such that the distance $m\delta_{max}$ is of a 'meaningful magnitude', how to ensure that the area $1 - \beta$ is exactly e.g. 80% at a given level of significance? The 'distance' between the two distributions in fig. 2.9.1 can be controlled by changing the number of elements n in each distribution. The

reason for this is that the variance of each of the distributions is given by $\frac{\sigma^2}{n}$. As n increases, the variance will decrease and the distributions will drift apart. Therefore:

$$\begin{aligned} \mu_{exp} + z_{a/2}\sigma/\sqrt{n} &= \bar{x}_{max} + z_{1-\beta}\sigma/\sqrt{n} \Leftrightarrow \\ \mu_{exp} - \bar{x}_{max} &= z_{1-\beta}\sigma/\sqrt{n} - z_{a/2}\sigma/\sqrt{n} \Leftrightarrow \\ m\delta_{max} &= \frac{\sigma}{\sqrt{n}}(z_{1-\beta} - z_{a/2}) \Leftrightarrow \\ \sqrt{n} &= \frac{\sigma(z_{1-\beta} - z_{a/2})}{m\delta_{max}} \end{aligned}$$

Squaring both sides of the equation and recalling that $-z_{a/2} = z_{1-a/2}$, yields:

$$n = \frac{\sigma^2(z_{1-\beta} + z_{1-a/2})^2}{m^2\delta_{max}^2} \tag{2.31}$$

However this assumes that the number of elements in each distribution is the same. When *SigniSite* analysis is performed, this is only the case, when $n_{aa} = \frac{N}{2}$. Due to the before mentioned symmetry, it does not make a difference whether we look at $n_{aa} > N - n$ or $n_{aa} < N - n$ (see fig. 2.7.1). In the following we will let $n_1 = n_{aa}$ be the smaller sample size and $n_2 = N - n$ the larger. This discrepancy in sample sizes can be corrected by introducing a ratio k, such that $k = \frac{n_2}{n_1}$, where $n_1 < n_2 \Leftrightarrow n < N - n \Leftrightarrow n < \frac{N}{2}$, we get:

$$k = \frac{n_2}{n_1} = \frac{N - n}{n} \tag{2.32}$$

If we assume that the variances are similar, except for the ratio k, we can calculate the estimated variance for both, by letting $\sigma_1 = \sigma_2$ and calculating the sum:

$$\sigma_1^2 + \frac{\sigma_2^2}{k} = \sigma^2 + \frac{\sigma^2}{k} = \frac{k\sigma^2}{k} + \frac{\sigma^2}{k} = \frac{\sigma^2(k+1)}{k} = \frac{k+1}{k} \cdot \sigma^2 \tag{2.33}$$

The calculated corrections for difference in elements, can be combined with eqn.

2.31, whereby:

$$n = \frac{k+1}{k} \cdot \frac{\sigma^2 \cdot z^2}{m^2 \delta_{max}^2} \tag{2.34}$$

Where $z = z_{1-\beta} + z_{1-\alpha/2}$ (equivalent with eqn. 8.27 in [127]). Knowing that:

$$\sigma = \sqrt{\frac{(N-n)(N+1)}{12n}} \tag{2.35}$$

and:

$$\delta_{max} = |\mu_{exp} - \bar{x}_{max}| = \left| \frac{N+1}{2} - \frac{n+1}{2} \right| = \frac{N-n}{2} \tag{2.36}$$

We can solve eqn. 2.34 for n, yielding 4 solutions:

$$\frac{1}{2}N \pm \sqrt{\frac{1}{4}N^2 \pm \frac{z}{3m}\sqrt{3N^2+3N}} \tag{2.37}$$

However given constraints $0 < n < N$, where $n, N \in \mathbb{N}$, 2 solutions can be excluded and we get:

$$n_{lower} = \frac{1}{2}N - \sqrt{\frac{1}{4}N^2 - \frac{z}{3m}\sqrt{3N^2+3N}} \tag{2.38}$$

$$n_{upper} = \frac{1}{2}N + \sqrt{\frac{1}{4}N^2 - \frac{z}{3m}\sqrt{3N^2+3N}} \tag{2.39}$$

Thus if a given residue *aa* at a given position p_i in an alignment of N sequences to be included in the set of tests, then $n_{lower} < n_{aa} < n_{upper}$. Once again as expected the solutions are symmetrical, such that $n_{upper} = N - n_{lower}$, in other words $t < n_{aa} < N - t$, where $t = n_{lower}$. $z = z_{1-\beta} + z_{1-\alpha/2}$, where standard values are $z = 0.84 + 1.96 = 2.8$. Recall that $m \in]0, 1]$ is the fraction of the distance between the expected mean rank and the maximum obtainable, which is not to be overlooked, i.e. $\delta_{target} = m\delta_{max}$. Example: Given an alignment of $N = 50$ sequences, a level of significance of 95%, a power of 80% and $m = 0.5$, the

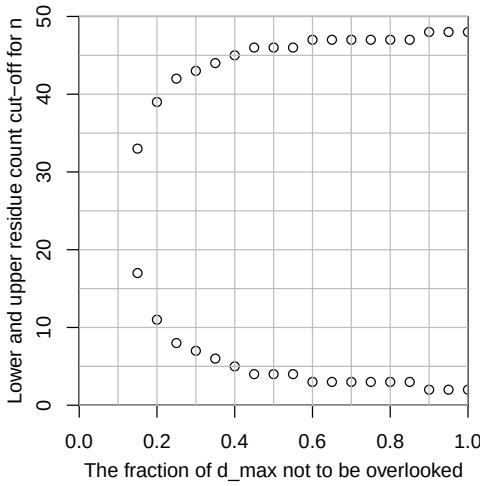


Figure 2.9.2: Power based residue count filtering. Depiction of the solutions given in eqn. 2.38 and eqn. 2.39. On the x-axis is $m \in]0, 1]$ in steps of 0.05. On the y-axis is n_{min} and n_{max}. $N = 50$, level of significance is $95\% \Rightarrow z_{1-\alpha/2} = 1.96$ and power is $80\% \Rightarrow z_{1-\beta} = 0.84$.

residue count cut-off t is calculated as:

$$\begin{aligned} t &= \frac{1}{2}N - \sqrt{\frac{1}{4}N^2 - \frac{z}{3m}\sqrt{3N^2 + 3N}} \\ &= \frac{1}{2}50 - \sqrt{\frac{1}{4}50^2 - \frac{2.8}{3 \cdot 0.5}\sqrt{3 \cdot 50^2 + 3 \cdot 50}} \\ &= 3.5 \end{aligned}$$

Since $n \in \mathbb{N}$, this corresponds to $n_{lower} = 4$ and $n_{upper} = 46$. Fig. 2.9.2 depicts this calculation for different values of m. Fig. 2.9.2 illustrates how a larger number of residues n_{aa} is needed if one wishes to be 80% of detecting a gradually smaller difference between $\overline{x}_{obs}$ and μ_{exp}, compared to $\overline{x}_{max}$ $(m = 1.0)$.

2.10 Filtering on a desired number of tests

An alternative approach could be to filter on a desired number of tests, rather than the number of a given residue. One might set an upper limit of 100 tests,

which are to be performed and then start with the residues with the best prerequisites, i.e. the ones which are at the theoretical optimal conditions, i.e. $n_{aa} = \frac{N}{2}$. Once the number of these residues have been computed, the next step is to include the residues with the second best theoretical optimal conditions, i.e. $n_{aa} = \frac{N}{2} - 1$ and $n_{aa} = \frac{N}{2} + 1$. This way more and more residues can be included based on their count distance to theoretical optimal conditions:

$$\frac{N}{2} - t \leq n_{aa} \leq \frac{N}{2} + t \tag{2.40}$$

Where $t \in \mathbb{N}$ is expanded in a step interval around $\frac{N}{2}$, thus including more and more residues until a certain number of tests are reached. This way one would be able to choose only to perform the before mentioned 100 tests and the step t where the number of tests exceed the limit for the first time can be calculated. Note that this does not mean that they *will* obtain a high *z-score*, the residues could easily be uniformly distributed over the entire p_i, it just means that, given their count, they have the *potential* to obtain a high *z-score*.

2.11 Summary

By analysing the *SigniSite* framework, we propose the following 3 methods for reducing the total number of tests performed. Filtering on positional residue count n_{aa} based on:

1. N and $z_{adj}(\alpha, n_t)$
   $t < n_{aa} < N - t, t = \frac{1}{2}N - \sqrt{\frac{1}{4}N^2 - \frac{1}{3}\left[z_{adj}(\alpha, n_t)\right]^2 (N+1)}$
2. N and power $(z_{1-\beta} + z_{1-\alpha/2})$
   $t < n_{aa} < N - t, t = \frac{1}{2}N - \sqrt{\frac{1}{4}N^2 - \frac{1}{3m}\left[z_{1-\beta} + z_{1-\alpha/2}\right]\sqrt{3N(N+1)}}$
3. n_t
   $\frac{N}{2} - t \leq n_{aa} \leq \frac{N}{2} + t, 0 \leq t \in \mathbb{N} \leq \frac{N}{2}$

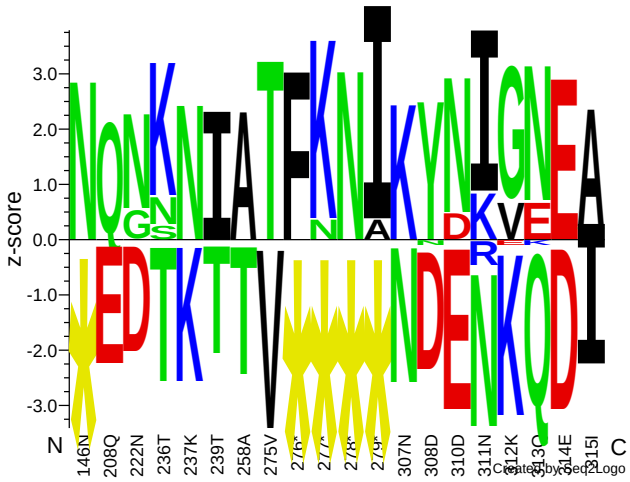


Figure 2.12.1: VAR2CSA-DBL5ε-infant birth weight sequence logo. 56 VAR2CSA-DBL5ε sequences sorted on infant birth weight was submitted to *SigniSite* with default settings, except for no correction for multiple testing. The sequence logo quantifies the strength of the residue associations to the infant birth weight. Amino acid residues, for which $y > 0$ are associated with a high infant birth weight and residues for which $y < 0$, with low infant birth weight. Residues with a *SigniSite z-score* larger than zero and are predominantly found among the top ranks of the sorted aligned VAR2CSA-DBL5ε sequences. The amino acids are colored according to their chemical properties as follows: Acidic [DE]: red, Basic [HKR]: blue, Hydrophobic [ACFILMPVW]: black and Neutral [GNQSTY]: green. [78]

2.12 Results

In order to apply the proposed methods for reducing the total number of tests performed, when analysing inhomogeneous systems, we turned to a multiple alignment of $N = 56$ sequences from the analysis of VAR2CSA-DBL5ε [61]. The phenotype of this particular alignment is the recorded birthweight of the infant. Submitting to *SigniSite* with default settings ($a = 0.05$, CMT using 'Bonferroni single-step'), no residues can be identified as significantly associated with the dataset phenotype. Re-submitting with no CMT, produces the logo in fig. 2.12.1. The motif '`TFKNI`' at $p275 - 279$ was predicted in the analysis of VAR2CSA-DBL5ε [61] and subsequently confirmed using the HDPMa

technique (VAR2CSA 3D7, Sector 2, 2263-GMDEFKNTFKNIKE-2276). This motif will therefore be used as true positive in the following. Tables 2.12.1 and 2.12.2 summarises the residue count and computed *z-scores* for the 'TFKNI'-motif. So

p_i	A	N	I	K	F	T	V	-
275	0	0	0	0	0	13	43	0
276	0	0	0	0	26	0	0	30
277	0	13	0	13	0	0	0	30
278	0	26	0	0	0	0	0	30
279	13	0	13	0	0	0	0	30

Table 2.12.1: VAR2CSA-DBL5ε TFKNI-motif residue counts. p_i refers to the position in VAR2CSA. The header are one-letter codes for amino acid residues, where '-' denotes a gap.

p_i	A	N	I	K	F	T	V	-
275	0	0	0	0	0	3.41	-3.41	0
276	0	0	0	0	3.21	0	0	-3.21
277	0	0.38	0	3.41	0	0	0	-3.21
278	0	3.21	0	0	0	0	0	-3.21
279	0.38	0	3.41	0	0	0	0	-3.21

Table 2.12.2: VAR2CSA-DBL5ε TFKNI *SigniSite z-scores*. p_i refers to the position in VAR2CSA. The header are one-letter codes for amino acid residues, where '-' denotes a gap.

the question is if any of the 3 proposed methods are able of reducing the number of tests sufficiently for us to identify the motif 'TFKNI' as significantly associated with the birth weight?

Calculating the sum of all unique amino acid resides at each position, we get $n_t = 533$ ($z_{adj} = 3.91$), from this we subtract all the fully conserved positions and we get $n_t = 347$ ($z_{adj} = 3.80$). If we then apply proposed method 1, we get $6 \leq n_{aa} \leq 50$, further reducing the number of tests to $n_t = 252$ ($z_{adj} = 3.72$). Proposed method 2 applied, at $\alpha = 0.05 \Rightarrow z_{1-\alpha/2} = 1.96$, $\beta = 0.20 \Rightarrow z_{1-\beta} = 0.84$ and $m = 0.25$ yields $8 \leq n_{aa} \leq 48$, further reducing the number of tests to $n_t = 227$ ($z_{adj} = 3.69$). Next we apply proposed method 3

and choose to perform $n_t = 100$ tests. This gives a total of $n_t = 101$ tests performed, when t=10, i.e. $18 \leq n_{aa} \leq 38$ ($z_{adj} = 3.48$). If we lastly apply proposed method 3 and choose to perform 50 tests, we get $n_t = 52$ tests ($z_{adj} = 3.30$) performed, when t=5, i.e. $23 \leq n_{aa} \leq 33$. Table 2.12.3 summarises these results:

Method	n_{min}	n_{max}	n_t	z_{adj}	$\Delta(n_t)[\%]$	$\Delta(z_{adj})[\%]$
0	1	55	347	3.80	0	0
1	6	50	252	3.72	-27.4	-2.1
2	8	48	227	3.69	-34.6	-2.9
3a	18	38	101	3.48	-70.9	-8.4
3b	23	33	52	3.30	-85.0	-13.2

Table 2.12.3: Reduction of tests using 3 proposed methods. Prerequisites where applicable are $\alpha = 0.05 \Rightarrow z_{1-\alpha/2} = 1.96$, $\beta = 0.20 \Rightarrow z_{1-\beta} = 0.84$, 'Bonferroni single-step' method for correcting for multiple testing, $m = 0.25$. m denotes the fraction of the distance $\delta_{max} = m|\mu_{exp} - \bar{x}_{max}|$, which we do not wish to overlook, with a power of $1 - \beta$.

If we compare table 2.12.3 with tables 2.12.1 and 2.12.2, we see that only when using method '3b' are we able to lower the significance threshold, such that z_{max}('TFKNI')$> z_{adj}$, however this limits n_{aa}, such that only residues for which $23 \leq n_{aa} \leq 33$ are included in the set of tested residues. This however excludes 'T-K-I', leaving only '-F-N-' (see table 2.12.1).

2.13 Evaluating MSA size required for *SigniSite* analysis

We can calculate how many sequences are required to obtain a significant result under optimal conditions, i.e. $N(z)$, by combining eqn. 2.21 and eqn. 2.19:

$$\begin{aligned} z_{max} &= \sqrt{\frac{3n(N-n_{aa})}{N+1}} \\ &= \sqrt{\frac{3\frac{N}{2}(N-\frac{N}{2})}{N+1}} \\ &= \frac{N}{\sqrt{\frac{4}{3}(N+1)}} \end{aligned}$$

This gives us an expression for the maximum obtainable *z-score* as a function of n, but since we are interested in evaluating the inverse, we solve the equation for z, yielding:

$$0 = -N(z)^2 + \frac{4}{3}z^2N(z) + \frac{4}{3}z^2 \tag{2.41}$$

Using CAS tool to solve the equation yields:

$$N(z) = \frac{2}{3}z\left(z + \sqrt{z^2+3}\right) \tag{2.42}$$

Setting $\alpha = 0.05 \Rightarrow z_{1-0.05/2} = 1.96$ and rounding up to nearest integer yields $N_{min}(1.96) = 6$. So if only 1 test is performed or no correction is applied a minimum of 6 sequences are required to obtain $z > z_{adj}$. Table 2.13.1 gives a short overview of the impact of performing $n = (1, 100, 1000)$ tests. Second to last column is the number of sequences required under optimal conditions, i.e. $rank = \frac{n_{aa}+1}{2}$ and $N = \frac{N}{2}$, however biology rarely offer optimal conditions for mathematical systems. The last column therefore are the minimum number of sequences, when the rank is twice the optimum and only half the number of optimal residues are present, i.e. $rank = 2 \cdot \frac{n+1}{2}$ and $n_{aa} = \frac{1}{2} \cdot \frac{N}{2}$:

n	α	z_{adj}	N_{min}^{opt}	N_{min}
1	0.05	1.96	6	20
100	0.0005	3.48	18	54
1,000	0.00005	4.06	23	71

Table 2.13.1: Threshold impact of the number of tests.

2.14 Discussion

Here we have demonstrated how it is possible to reduce the total number of tests performed in a system prior to analysis, based on intrinsic properties of the *SigniSite* framework. So why is that despite reducing the number of tests with as much as 85%, we are only barely able to identify two of the five residues in the VAR2CSA TFKNI motif? Looking at fig. 2.4.1B it becomes evident that when correcting the α-value with the number of tests, the resulting depiction of $(n_t, z_{adj}(n_t))$ plateaus relatively fast. The consequence of this is, that once the number of tests are in a 'plateau' range, it does not affect the adjusted *z-score* threshold if e.g. 100 tests can be excluded. It is simply a property of the non-linearity of the CMT.

Also it is important to realise that sometimes, you may have something which is a true positive, but due to lack of data and thereby lack of ability to sufficiently separate the H_0 and H_1 hypotheses, you are not able to identify the true positive and it therefore remains a false negative. Basically given enough data, you can 'prove' anything, given to little data, you can prove 'nothing'.

Excluding entire MSA positions from analysis using information content is problematic in that, as previously shown, the strongest significance is when $f = 0.5$. Using e.g. the Shannon Entropy and choosing a conservation cut-off of conventionally 1 is thus difficult in that given a position with 2 residues, $f_1 = 0.5$ and $f_2 = 0.5$, then the Shannon entropy will be 1 and if $f_1 = 0.49$ and $f_2 = 0.51$, then the Shannon entropy would be less than 1. If the most variable positions are sought, positions with a 50/50 distribution will be considered conserved and thus left out. Furthermore filtering is applied per position for variability versus

per residue for power calculation. Thus risking to exclude single residues based on the composition of the rest of the residues at the position.

So all 3 methods will allow parameter adjustment. However to be statistically correct, one should always decide on a set of parameters of choice, run the analysis and accept the results, otherwise we will slowly but steadily be embarking on the dangerous path of *p-value* hacking.

2.15 *SigniSite* analysis of MHCI:peptide binding complex

2.15.1 Introduction

Given the benchmark in [78], where *SigniSite* clearly outperformed competing state-of-the-art methods, we found it interesting to apply the method to an in-house data set of 9-mer peptides with measured binding affinities to a set of MHC-I HLA-A and HLA-B alleles.

2.15.2 Materials

Tables 2.15.1 and 2.15.2 summarises the data set submitted to *SigniSite analysis*.

Number of	HLA-A	HLA-B
Measurements	76,716	52,256
Unique measurements	29,351	13,291
Unique peptides	16,998	11,094
Unique alleles	42	49

Table 2.15.1: Overview of MHCI:peptide data.

Combining the two HLA-A and HLA-B datasets, yielded a total of 128,972 datapoints. This is the dataset used in the well proven NetMHC-3.0 [100], currently in version 3.4. Tables 2.15.1 and 2.15.2 illustrates that the data is biased towards strong binders and also fairly tied, with respect to the values of the

Description	HLA-A	HLA-B
Minimum	0.00	0.00
Maximum	1.00	1.00
Mean	0.26	0.20
Standard deviation	0.29	0.25
Median	0.08	0.08

Table 2.15.2: Overview of MHCI:peptide stats.

measurements. This is clear, when comparing the total number of measurements, with the number of unique measurements.

2.15.3 Methods

As a proof of concept, we identified the number of *different* peptides, with measured binding affinity to the *same* HLA-A/B allele. Then we took the HLA-A and HLA-B allele with most measurements, i.e. HLA-A0201 (9,120) and HLA-B1501 (4,214). This way we could create two *SigniSite* compatible alignments and compute the *SigniSite* position-specific scoring matrices (PSSM). These could then be correlated with the last PSSM for the same alleles computed by the MHC motif viewer [119] using the non-parametric Spearman's Correlation Coefficient. (non-parametric since in this context it is not interesting whether the exact same values are assigned, but rather the decision on which position and residues are the most important).

SigniSite compatible datasets were compiled, by identifying peptides, with associated binding affinity to multiple HLA-A and HLA-B alleles. MSAs was then created, each containing the set of *different* HLA-A and HLA-B α_1 domain sequences ($p_1 - p_{180}$), which all had been measured to the *same* peptide. This way the two largest MSAs contained 43 HLA-A/B sequences. In order to address the previous mentioned problem of ties in the dataset, we only included the MSAs which contained at least $N_{seqs}/2$ (rounding odd numbers down) unique associated values. Lastly the minimum number of sequences in the MSAs was set to 20. This way, the final data set consisted of 415 MSAs, which were submitted

to *SigniSite* analysis and the resulting PSSMs were stored.

In order to evaluate the predictive performance of *SigniSite* based on the computed PSSMs, we defined specificity determining positions as those MHCI a_1 positions, which was found to be in contact with the 9-mer peptide within a distance of 4 angstroms, based on a structural analysis of HLA structures available in the PDB as defined in [100]. This yielded a list of 34 actual positives and 146 actual negatives. These 34 actual positives is referred to as the *'pseudo-sequence'* of the MHCI a_1. Per position we then transformed the SigniSite z-scores to positional predictions, by computing the maximum of the absolute z-scores, $|z|_{max}(p_i)$, and the sum of the absolute z-scores, $\sum_{p_i} |z|$. These measures were then compared with the Shannon Entropy and the Kullback-Leibler Divergence.

Furthermore a meta analysis was performed aiming at identifying MHC-I positions, which were consistently identified as binding determinants. A data set was constructed by selecting all HLA-A/B MSAs, containing at least 20 sequences and for each MSA the number of unique associated values must be at least half the number of sequences. This resulted in a total of 415 MSAs. Foreach of these MSA, $\mathrm{H}(p)$ and $\sum_{p_i} |z|$ was calculated and ranked, such that the result of each calculation was a single column, where the highest values was assigned a rank of 1. Each column was added to a matrix, such that the final matrix had a dimension of 180 rows and 415 columns. Foreach row in this matrix, the mean of the ranks was calculated and the resulting vector was once again ranked, assigned rank 1 to the lowest value. This way we could not only assess the degree of agreement between the $\mathrm{H}(p)$ and $\sum_{p_i} |z|$ scores, but also intersect the ranked mean of ranks with the list of known contact positions.

2.15.4 Results

Correlating the *SigniSite* PSSM with that of the MHC motif viewer yielded $\mathrm{SCC}(\mathrm{A}^*02{:}01) = 0.845$ and $\mathrm{SCC}(\mathrm{B}^*15{:}01) = 0.804$.

Fig. 2.15.1 show computed AUC values for the 415 MSA containing $N = 20$ or more HLA-A/B sequences and at least $N/2$ unique associated values. Target

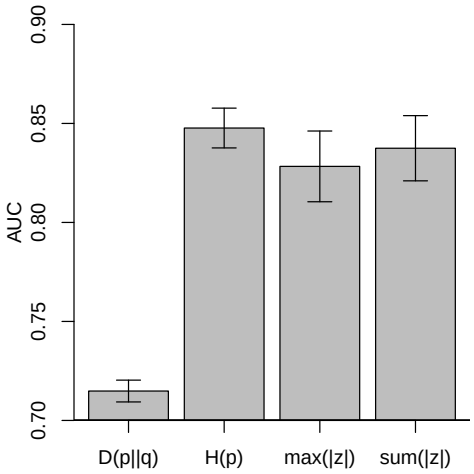


Figure 2.15.1: Graphical representation of the performance of *SigniSite*. Overview of AUC values. $D(p||q)$ is the Kullback-Leibler divergence, $\mathrm{H}(p)$ is the Shannon Entropy and $max(|z|)$ is the positional absolute maximum *SigniSite z-score* and lastly $sum(|z|)$ is the positional absolute sum hereof. The AUC values were computed based on the 4 methods of scoring each of the 180 positions in the MHC-I α_1 domain against a list of 34 positional true positives and 146 true negatives defined as residues being within a distance of 4 angstrom to the 9-mer peptide in the binding groove [100].

positions were assigned based on MHCI residues potentially in contact with the 9-mer peptide in the binding groove within a distance of 4 angstrom, as defined in [100]. *SigniSite z-scores* were transformed to positional scores by assigning maximum absolute *z-score* $(max(|z|))$ and positional sum of absolute *z-scores* $(\sum |z|)$. Figure 2.15.1 depicts the means and standard deviation for the performance measured using the measure of Area under the Receiver operating characteristic (ROC) curve (AUC) [111]. Values are:
$D(p||q) = 0.7148 \pm 0.0055$, $\mathrm{H}(p) = 0.8477 \pm 0.01$,
$max(|z|) = 0.8283 \pm 0.0179$ and $sum(|z|) = 0.8375 \pm 0.0165$.

Table 2.15.3 summarises results for Welch Two Sample t-test of *SigniSite* performance.

Fig. 2.15.2 depict the results of the meta-rank analysis. Details in figure legend.

	$D(p\|\|q)$	$\mathrm{H}(p)$	$max(\|z\|)$	$sum(\|z\|)$
$D(p\|\|q)$		<2.2e-16	<2.2e-16	<2.2e-16
$\mathrm{H}(p)$			<2.2e-16	<2.2e-16
$max(\|z\|)$				4.232e-14
$sum(\|z\|)$				

Table 2.15.3: Welch Two Sample t-test of *SigniSite* performance.

Both scoring methods were in agreement regarding the ranking of the positions, with relation to impact on MHCI:peptide binding $SCC = 0.996$.

Fig. 2.15.3 maps MHCI a_1 positions 9, 45, 62, 66, 67, 70, 95, 97, 114, 116, 156 and 163, which are consistently top-ranked by *SigniSite* and the Shannon entropy, mapped onto HLA-A*02:01, PDB-ID 1i4f [71].

Fig. 2.15.3 depict how the amino acid residues of the top-ranked positions have side-chains facing the the binding groove.

2.15.5 Discussion

The transformed *SigniSite* scores outperforms Kullback-Leibler Divergence, but not the Shannon Entropy. When comparing to simple Shannon entropy, it seems we fall victims to Ockham's razor. However *SigniSite* offers positional residue resolution, in fact *SigniSite* assigns *z-scores* to each unique residue at each position and in order to be compared with the Shannon entropy, we must transform the *SigniSite z-scores*. It can be argued, that we are comparing two methods, which really are not comparable.

Computational analysis of binding interactions reveal that the sites denoted as contact sites are distributed evenly across the meta-ranks. Fig. 2.15.3 show the side-chains of the amino acid residues of the top-ranked positions have side-chains facing the binding groove. By visual inspection of the HLA-A*02:01 structure it seems likely that the identified position does indeed constitute the primary binding determinant. Based in this, it could be interesting to re-evaluate which positions should define the MHC-I pseudo sequence. This could be done

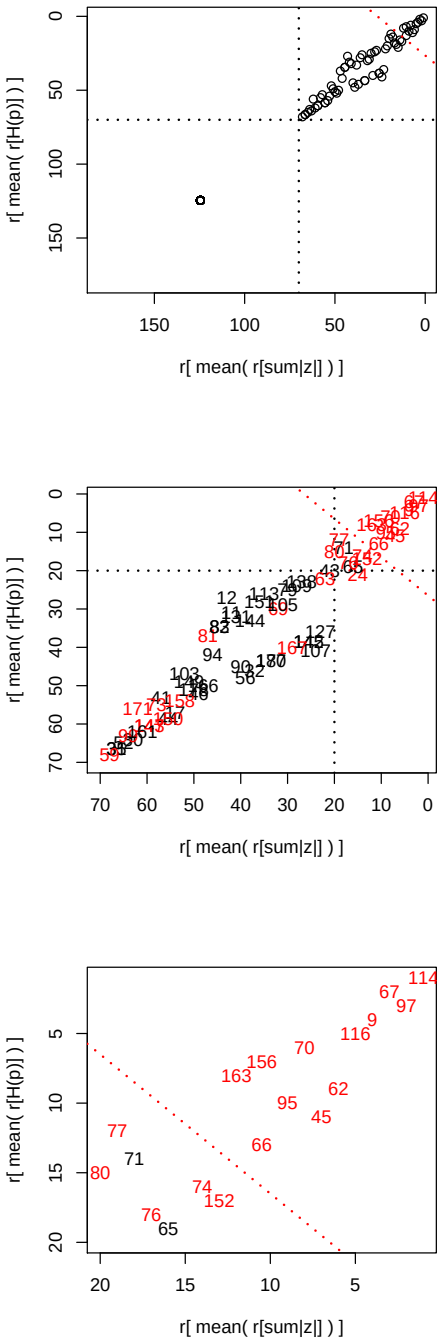


Figure 2.15.2: Meta-ranking of MHC-I α_1 positions. On the axes are the meta-ranks computed using x, the positional *SigniSite* $\sum |z|$ and y, the positional Shannon Entropy. (Note the axes have been reversed). The numbers in the plot represent each of the 180 positions of the MHC α_1 domain. If a number is in red font, it indicated that it is part of the pseudo-sequence, i.e. the set of 34 residues known to be in physical contact with the bound peptide. Positions in black font, the position is not part of the pseudo-sequence. From top plot to bottom plot, each plot is zoomed in compared to the prior. **Top**: The data-point in the lower left corner represent the fully conserved positions, where both the *SigniSite z-score* and the Shannon Entropy have zero values. The red dotted line seem to separate the positions with a consistent high rank, from what could be less impacting positions. **Middle**: Zooming further in, it becomes evident that both positional scoring methods assign low ranks to positions in the pseudo-sequence. **Bottom**: Positions 9, 45, 62, 66, 67, 70, 95, 97, 114, 116, 156 and 163 are consistently top-ranked by both methods.

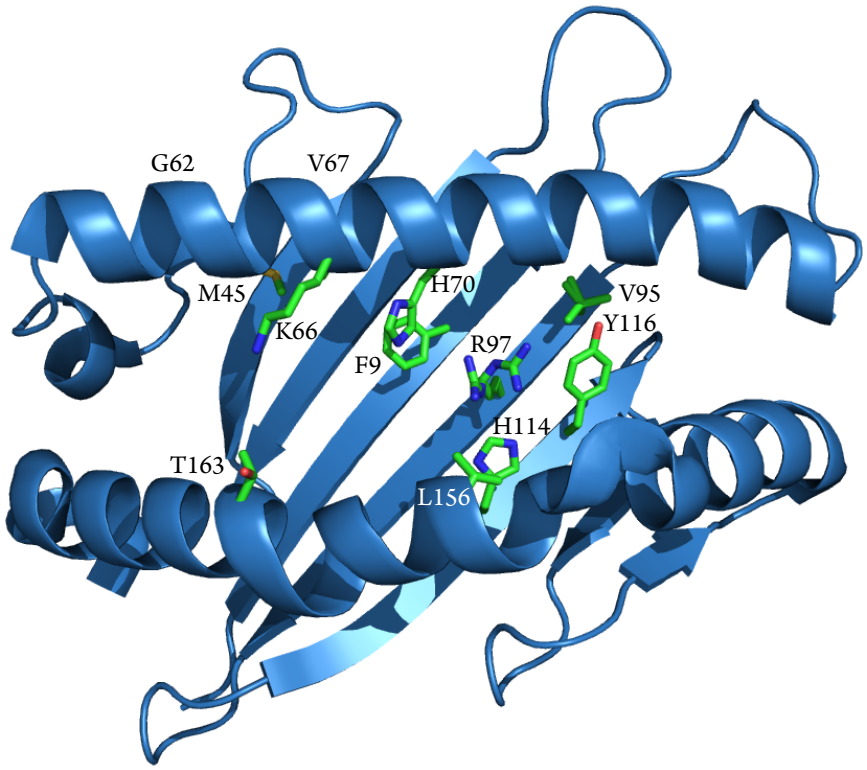


Figure 2.15.3: HLA-A*02:01 mapping of top meta-ranked α_1 positions 9, 45, 62, 66, 67, 70, 95, 97, 114, 116, 156 and 163. Positions and one-letter coded amino acid residues relative to HLA-A*02:01 are given beside the side-chains of the top-ranked positions.

by redefining the pseudo-sequence according to the top-ranked positions and then re-train the NetMHC-3.0 method to see if any difference in performance can be obtained.

2.16 Paper I

The following paper was published in Nucleic Acids Research in July 2013. The paper focuses on the use of the web server available at http://www.cbs.dtu.dk/services/SigniSite/. For details on the *SigniSite* method, please see the supplementary materials.

W286–W291 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* *Published online 12 June 2013*
doi:10.1093/nar/gkt497

SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments

Leon Eyrich Jessen[1], Ilka Hoof[2], Ole Lund[1] and Morten Nielsen[1,3,*]

[1]Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark, [2]Department of Molecular Biology and Biotech Research and Innovation Centre (BRIC), Bioinformatics Centre, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen, Denmark and [3]Instituto de Investigaciones Biotecnologicas, Universidad Nacional de San Martin, San Martin, B 1650 HMP, Buenos Aires, Argentina

Received January 31, 2013; Revised May 2, 2013; Accepted May 15, 2013

ABSTRACT

Identifying which mutation(s) within a given genotype is responsible for an observable phenotype is important in many aspects of molecular biology. Here, we present *SigniSite*, an online application for subgroup-free residue-level genotype–phenotype correlation. In contrast to similar methods, *SigniSite* does not require any pre-definition of subgroups or binary classification. Input is a set of protein sequences where each sequence has an associated real number, quantifying a given phenotype. *SigniSite* will then identify which amino acid residues are significantly associated with the data set phenotype. As output, *SigniSite* displays a sequence logo, depicting the strength of the phenotype association of each residue and a heat-map identifying 'hot' or 'cold' regions. *SigniSite* was benchmarked against SPEER, a state-of-the-art method for the prediction of specificity determining positions (SDP) using a set of human immunodeficiency virus protease-inhibitor genotype–phenotype data and corresponding resistance mutation scores from the Stanford University HIV Drug Resistance Database, and a data set of protein families with experimentally annotated SDPs. For both data sets, *SigniSite* was found to outperform SPEER. *SigniSite* is available at: http://www.cbs.dtu.dk/services/SigniSite/.

INTRODUCTION

Whether conducting research in vaccine design or trying to elucidate the intimate details of a given receptor::ligand interaction, genotype–phenotype correlation is a powerful tool to enhance the understanding of the minute subtleties, often characterizing research within the field of molecular biology.

The traditional approach for wet-laboratory analysis of genotype–phenotype correlations involves site-directed mutagenesis and subsequent quantification of mutation-impact on the phenotype, e.g. binding-affinity or catalytic efficiency. This approach of mutating all amino acid residues in a given protein is a time consuming and tedious task. Random mutagenesis has the advantage of introducing a large number of random mutations throughout the protein. One example of application of random mutagenesis is to increase the signal from near-infrared fluorescent proteins (1). In such a panel of sequenced variants with multiple mutations, it is a complex task to systematically pinpoint the exact amino acid residue(s), i.e. the genotype, associated with a given phenotype (e.g. fluorescence). Another area of application is genotype–phenotype association studies in proteins, which show inherent natural variability, as is the case for instance for proteins involved in the pathogenesis of malaria (2).

Here, we present *SigniSite*, an online application for subgroup-free residue-level genotype–phenotype correlation in protein multiple sequence alignments (MSAs). A number of methods have been developed for the identification of functional sites in protein sequences (3–10), most requiring a definition of functional subgroups before analysis. However, if the phenotype associated with the sequences is not categorical (e.g. substrate-specificity) but continuous (e.g. catalytic efficiency), a pre-division of sequences subgroups is none trivial. In contrast, *SigniSite* does not require any subgroup division or binary classification. Instead, *SigniSite* directly analyses the raw sequences and associated continuous values. The main novelty of *SigniSite* is that unlike conventional methods for the prediction of specificity determining

*To whom correspondence should be addressed. Tel: +45 45 25 61 27; Fax: +45 45 93 15 85; Email: jessen@cbs.dtu.dk

positions (SDP), it not only predicts the positions in the MSA determining a given protein function but also makes a statistical evaluation of which types of amino acid residue substitutions (genotype) are associated with the observable phenotype at the SDP.

The web server implementation of the *SigniSite* method described here is an automatized online application with an easy-to-interpret graphical output. The application is easy to use for the non-expert end-user and aims at aiding researchers in the analysis of sequence data, where the phenotype is quantified by a real number. A list of abbreviations is available in the Supplementary Data.

THE WEB SERVER

User interface

The *SigniSite* server is intended to provide the non-expert user with a simple interface. At default settings, an amino acid residue is considered significantly associated with the MSA phenotype, if the *P*-value for the specific residue is smaller than or equal to $\alpha = 0.05$ after Bonferroni Single-Step Correction for Multiple Testing (CMT) (11). On the submission page, sequences can be submitted to the server either as paste-in or via the file upload field. On submission, *SigniSite* will check whether the submitted sequences are aligned. If not, an MSA will be created using MAFFT (12). *SigniSite* will exclude any characters other than the one-letter representation of the 20 standard proteogenic amino acids from the analysis.

Input

As input *SigniSite* takes an MSA in FASTA-format (minimum two sequences). Each sequence must have an associated real number, stated white-space-separated as the last element in its FASTA header. At least two different values must exist in the MSA. The MSA is assumed pre-sorted, if the end-placed value is absent. A section with options for customizing the analysis is available. The following parameters are user-adjustable: (i) the level of significance 'α', $0 \leq \alpha \leq 1$ (default is 0.05). (ii) The method for CMT: 'Bonferroni Single-Step' (default), 'Holm Step-Down' (11) or 'no correction'. (iii) The sorting of the sequences: 'Decreasing', highest sequence-associated value is considered the strongest, e.g. fluorescent protein signals, and vice versa for 'Increasing', e.g. binding affinity. Furthermore, the user can choose a reference sequence to assign sequence-specific positional output numbering. This is useful, when the MSA contains insertions. Finally, the user can modify the logo output by choosing to include either 'Significant positions' (default, displays all residues at positions where at least one amino acid residue has been identified as significantly associated with the data set phenotype), 'Significant Residues' (as for significant positions, but only including significant residues) or 'Full Logo' (all residues at all positions). At the results page, a button below the generated logo allows the user to fully customize the logo using Seq2Logo (13).

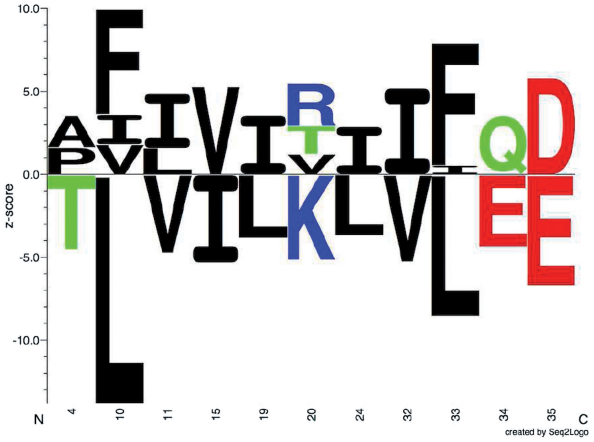


Figure 1. Sequence logo. Example of sequence logo (13) output from *SigniSite* from the analysis of the ATV ~Antivirogram multiple sequence alignment (MSA), truncated to $p_1 - p_{35}$ for the purpose of illustration (see 'Materials and Methods' section). The analysis was performed with default settings. On the *x*-axis are the MSA positions p and on the *y*-axis the *Z*-scores for each amino acid residue a ($z_{p,a}$). The height of each letter representing the residues is proportional to $z_{p,a}$, i.e. the strength of the statistical association between the residue and the data set-phenotype. Residues above the $Z = 0$ line have a $z_{p,a} > 0$, i.e. enhances the phenotype, whereas residues below the $Z = 0$ line have a $z_{p,a} < 0$, i.e. inhibits the phenotype, e.g. the presence of a certain residue with favourable chemical properties may enhance binding ($z_{p,a} > 0$), whereas a residue with unfavourable properties may inhibit binding ($z_{p,a} < 0$). Colour-coding: acidic [DE]: red, basic [HKR]: blue, hydrophobic [ACFILMPVW]: black and neutral [GNQSTY]: green (14).

Output

The *SigniSite* output is intended to provide the end-user with an easily interpretable graphical representation of the statistical evaluations performed by *SigniSite*. An example of a sequence logo (13) generated by *SigniSite* is shown in Figure 1. The logo gives an overview of residue associations. See Figure 1 legend for further details. *SigniSite* will also generate a heatmap (Figure 2). The heatmap is intended to give a graphic overview of 'hot' and 'cold' regions in the MSA, with respect to the data set phenotype. See Figure 2 legend for details.

RESULTS

As an initial performance evaluation, we chose to analyse 18 human immunodeficiency virus type 1 (HIV-1) MSAs compiled from the Stanford University HIV Drug Resistance Database (15,16) (HIVdb) using Spearman's rank correlation (SCC) to correlate the obtained *SigniSite* *Z*-scores ($z_{p,a}$ for each residue a at each position p) with the table of resistance mutation scores (RMS) also available from the HIVdb (see 'Materials and Methods' section), i.e. $\mathrm{SCC}(z_{p,a} \sim \mathrm{RMS})$. Results are given in Table 1.

As the SCC evaluation is threshold dependent, a threshold-independent performance evaluation was added using the area under the receiver operator

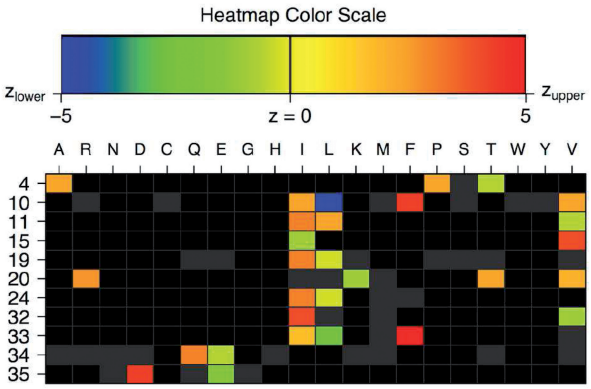


Figure 2. *SigniSite* heatmap from the analysis of the ATV ~Antivirogram multiple sequence alignment (MSA), truncated to p_1-p_{35} for the purpose of illustration (see 'Materials and Methods' section). The analysis was performed with default settings. On the *x*-axis are the 20 proteogenic amino acids *a* and on the *y*-axis the positions *p* in the analysed MSA. The colour coding of the fields is such that fields reflecting $z_{p,a} \leq -5$ are blue, whereas $z_{p,a} \geq 5$ results in a red field. For $-5 < z_{p,a} < 5$, nuances in between are used. If a residue has a $z_{p,a}$ of 0, the cell is coloured grey. Absent residues are coloured black. If only one grey cell is present at a given position, this implies that the position is fully conserved, harbouring only this residue. If more grey cells are present, their associated *P*-values have become $P = 1 \Rightarrow z_{p,a} = 0$ after correction for multiple testing.

Table 1. Benchmark results

Measure	$\|z\| \geq 0$	$\|z\| \geq 1.96$	$\|z\| \geq 1.96_{CMT}$
SCC[a]	0.451 ± 0.015	0.506 ± 0.016	0.542 ± 0.020
MCC[b]	0.492 ± 0.028	0.387 ± 0.027	0.297 ± 0.040
SENS[b]	0.915 ± 0.015	0.598 ± 0.056	0.386 ± 0.055
SPEC[b]	0.579 ± 0.016	0.774 ± 0.031	0.882 ± 0.022

[a]Calculated against the RMS.
[b]Calculated against the $(RMS+IAS)_{mut}$.
Measures are means ± SE. CMT: corrected for multiple testing, SCC: Spearman's rank correlation, MCC: Matthews Correlation Coefficient, SENS: sensitivity, SPEC: specificity.

characteristics curve (AUC) measure, resulting in $AUC(z_{p,a} \sim RMS_{bin}) = 0.791 \pm 0.010$. Certain mutations not included in the RMS were repeatedly identified by *SigniSite*. As the majority of these mutations were found in the binary resistance annotations from the international antiviral society-USA (IAS) (17), we enriched the RMS_{bin} with the IAS and re-calculated the AUC, obtaining a significant performance increase of $AUC(z_{p,a} \sim (RMS+IAS)_{mut}) = 0.822 \pm 0.011 (P = 5.16 \cdot 10^{-4})$, two-tailed paired *t*-test).

Furthermore, we evaluated the performance of *SigniSite* using performance measures: Matthew's correlation coefficient (MCC), sensitivity (SENS) and specificity (SPEC) against $(RMS+IAS)_{mut}$. See Table 1 for results.

Having obtained good results for both the threshold-dependent and -independent performance evaluations, we turned to benchmark *SigniSite* against similar existing methods. In a 2009 benchmark study (18), SPEER (5,19) was identified as the state-of-the-art

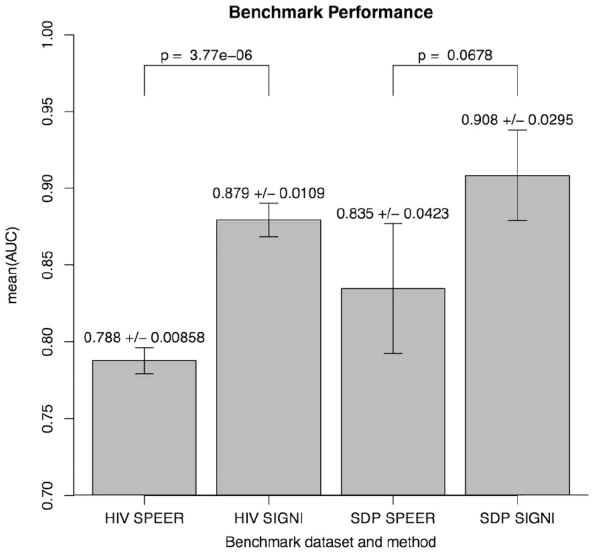


Figure 3. Measures are mean (AUC) ± SE. Columns are: HIV [SPEER/SIGNI], SPEER and *SigniSite's* predictions on the HIVdb data set. SDP [SPEER/SIGNI] SPEER and *SigniSite's* predictions on the SDP data set. *P*-values quantifying the significance of the difference in performance were obtained using a two-tailed paired *t*-test.

method for prediction of specificity definition positions (SDP). We, therefore, here compared the performances of *SigniSite* and SPEER on each of their original benchmarks data sets (see 'Materials and Methods' section) against $(RMS+IAS)_{pos}$. The results are shown in Figure 3. The results show that *SigniSite* outperforms SPEER on both data sets. The difference in predictive performance was, however, only found to be statistically significant for the HIVdb data set.

DISCUSSION

SigniSite aims at providing a simple-to-use method for subgroup-free residue-level genotype–phenotype correlation in protein MSAs. *SigniSite*, thus, addresses a long-existing challenge in molecular biology; genotype-phenotype mapping. Genotype–phenotype mapping has a wide range of purposes in molecular biology, e.g. structural regions responsible for immunity (2), identifying protein-variants responsible for the severity of a disease (20) or coupling receptor polymorphisms to surface expression (21) etc.

Site-directed mutagenesis in proteins and subsequent quantification of mutation-impact on a given phenotype is a time consuming and tedious task. High-throughput methods such as e.g. random mutagenesis (1) have, therefore, been developed. However, the challenge of analysing the increasingly larger volumes of data being generated only becomes greater. Additionally, large genotype–phenotype data sets (GPDs) can be compiled from publicly available databases, such as the HIVdb (15,16). *SigniSite* addresses this exact challenge.

SigniSite was benchmarked on publicly available GPDs and RMS from the Stanford University HIV Drug Resistance Database (HIVdb) (15,16). We observed that for each of the 18 different benchmark data sets, *SigniSite* consistently identified certain residues, not annotated in the RMS table, as significantly associated with anti-viral drug resistance. We compared these identifications with binary resistance annotations from the International Antiviral Society-USA (IAS) (17) and found that the majority were indeed annotated as resistance impacting. This observation suggests that the RMS data are not exhaustive, and that the obtained correlation should rather be regarded as a lower bound of the true predictive performance.

As the SDP method SPEER (5,19) was found to be the state-of-the-art method in a 2009 benchmark study (18), we chose to compare *SigniSite* to SPEER. We observed that *SigniSite* significantly outperformed SPEER on the HIVdb data set ($P = 3.77 \cdot 10^{-6}$) and for the SDP data set (as defined in the SPEER paper), *SigniSite* likewise outperformed SPEER, approaching a significant difference ($P = 0.0678$). Furthermore, *SigniSite* was much faster, taking only a few minutes to analyse the largest of the MSA ($n_{seqs} = 1{,}374$). SPEER on the other hand requires to be compiled in a slower version, when $n_{seqs} > 200$, taking $\sim$2 h to complete the analysis.

In conclusion, *SigniSite* provides two important novel features: (i) *SigniSite* does not require any manual annotation of the data before analysis, e.g. binder/non-binder classification, *SigniSite* requires only sequences and associated values. (ii) Unlike conventional SDP prediction methods like SPEER, *SigniSite* will not only identify positions impacting the phenotype but also pinpoint the exact amino acid residue substitution(s) responsible for the impact detected at the identified position. To the best of our knowledge, this level of resolution has so far not been available.

MATERIALS AND METHODS

Benchmark data sets

Summary, see Supplementary Data for details.

HIVdb resistance mutation scores

The table of RMS was downloaded from the HIVdb (15,16), available at http://hivdb.stanford.edu/DR/cgi-bin/rules_scores_hivdb.cgi?class=PI. The table of RMS contains information about positions known to harbour mutations ($n = 688$) compared with wild-type (WT) and their impact on resistance towards eight different protease inhibitors (PIs). Positive scores range is [3,60] ($n = 296$) and indicates that the mutation increases the resistance towards a given PI. Negative score range is $[-5, -10]$ ($n = 15$) and indicates a decreased resistance. Scores of 0 ($n = 377$) indicate lack of resistance impact. At each position annotated in the table of RMS, the consensus residue was assigned an RMS of 0.

IAS resistance annotations

Protease mutations known to impact PI resistance were retrieved from the table 'mutations in the protease gene associated with resistance to protease inhibitors', in the International Antiviral Society USA (IAS)'s Update of the Drug Resistance Mutations in HIV-1: March 2013 (17). Also here, the consensus residue at annotated resistance positions was assigned an IAS score of 0.

Table transformations

The following table transformations were performed: $\text{RMS} \rightarrow \text{RMS}_{bin}$, such that $\text{RMS} > 0 \Rightarrow \text{RMS}_{bin} = 1$, otherwise $\text{RMS}_{bin} = 0$. $\text{RMS}_{bin}+\text{IAS} \rightarrow (\text{RMS+IAS})_{mut}$, such that $\text{RMS}_{bin} > 0$ or $\text{IAS} > 0 \Rightarrow (\text{RMS+IAS})_{mut} = 1$, otherwise $(\text{RMS+IAS})_{mut} = 0$. $(\text{RMS+ IAS})_{pos}$, such that for each position in $(\text{RMS} + \text{IAS})_{mut}$ the resulting $(\text{RMS+IAS})_{pos} = 1$ if at least one $(\text{RMS+IAS})_{mut} > 0$, otherwise $(\text{RMS+IAS})_{pos} = 0$. In all tables, any score $s_{table} > 0$ is considered an actual positive and any score $s_{table} \leq 0$ is considered an actual negative (Table 2).

MSAs from the HIVdb protease GPDs

GPDs were downloaded from the Stanford University HIV Drug Resistance Database (HIVdb) (15,16) Version 5.0, March, 2012, available at http://HIVdb.stanford.edu/cgi-bin/GenoPhenoDS.cgi. MSAs were compiled from the GPDs. Each MSA contains the sequences of a set of HIV-1 protease variants with measured fold change in resistance (compared with WT) towards the same PI, measured using the same assay. Only PIs present in both the table of RMS and the GPDs were used limiting the analysis to 6 PIs: *ATV, IDV, LPV, NFV, SQV and TPV* each of which was assayed using the three assays: 'Antivirogram' (Virco™), 'PhenoSense' (ViroLogic™) and 'All Others'. A total of 12 714 sequences were constructed and compiled into 18 MSAs. The length of each of the protease variants is 99 amino acid residues.

The SPEER program and SDP benchmark data

SPEER, MSAs and corresponding experimentally annotated specificity determining sites were downloaded from the SPEER repository available at: ftp://ftp.ncbi.nih.gov/pub/SPEER/ (5,19). We downloaded the latest curated version of the data as described by Chakrabarti and Panchenko (18).

The *SigniSite* method

The method takes a set of (protein) sequences as input. If the sequences are not aligned, *Signisite* will use MAFFT (12) to make an MSA from the input sequences. Subsequently, the sequences are ranked with respect to a real number associated with each sequence, e.g. the replicative capacity or catalytic efficiency. For each amino acid at each position in the MSA, a non-parametric test is performed to test whether the observed ranks deviate significantly from the expected ranks. CMT of the resulting *P*-values may be performed using Bonferroni single-step or Holm step-down procedures. The resulting *Z*-scores per residue are visualized in a logo plot and a heatmap.

Table 2. Overview of target table notation

Notation	Format	Level	Annotating
RMS[a]	Real num.	Residue	Fold-change in PI resistance
IAS[b]	Binary	Residue	PI ass. resistance mutations
RMS_{bin}[c]	Binary	Residue	PI ass. resistance mutations
$(RMS+IAS)_{mut}$[d]	Binary	Residue	PI ass. resistance mutations
$(RMS+IAS)_{pos}$[e]	Binary	Position	Positions ass. with PI resistance

[a]It is used when calculating SCC, [b]it is used to look up mutations not annotated in **I**, but repeatedly identified by *SigniSite*, [c]it is used when calculating AUC, [d]it is used for the enriched AUC calculation and when calculating the MCC, SENS and SPEC, [e]it is used as positional targets, when comparing the predictive performances of *SigniSite* and SPEER.
'num.', 'ass.', 'PI' abbreviates 'numbers', 'association' and 'protease inhibitor'. In all tables, any score $s_{table} > 0$ is considered an actual positive and any score $s_{table} \leq 0$ is considered an actual negative.

Brief description of the method underlying SigniSite

(see Supplementary Data for details). Initially each sequence is assigned a rank by sorting the sequence associated values (either ascending or descending depending on type of value) and then assigning a rank of '1' to the first sequence after sorting, '2' to the second and so forth. Each amino acid residue a observed at position p ($res_{p,a}$) in the MSA is then assigned the rank of the sequence to which it belongs. This way each $res_{p,a}$ is associated with a specific rank. At each position in the MSA, the mean ranks of each residue type are then calculated and placed in a rank matrix, where each row corresponds to a position in the MSA and each column to one of the 20 standard proteogenic amino acids, sorted according to A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y and V (*SigniSite* will exclude any characters but these 20).

Subsequently, *SigniSite* evaluates for each position and residue type the difference between the mean of the observed and expected ranks. The mean of the expected ranks is the mean of the ranks we would observe if the residue type $res_{p,a}$ was randomly distributed over the column p in the MSA. This difference between observed and expected ranks is quantified by a Z-score assigned to each residue type at each position, yielding a Z-score-matrix. If a given position is fully conserved, $z = 0$ is assigned to the conserved residue. If a given residue type is absent at a given position, $z = 'NA'$ is assigned.

The non-parametric statistics, on which *SigniSite* is based, are similar to that of Wilcoxon test statistics (22), where the obtained evaluation scores can be approximated by the standard normal distribution, thus allowing Z-score conversion to P-values by standard method. As one test is performed per residue type, per position, *SigniSite* will by default apply Bonferroni single-step (11) CMT to adjust the reported P-values.

Benchmarking

For each of the 18 MSAs compiled from the HIVdb GPDs (see 'Materials and Methods' section), a set of predictions were made (Z-scores) estimating the strength of the association of each residue type a at each position p ($z_{p,a}$) to the phenotype of the MSA. The obtained set of $z_{p,a}$'s was then correlated with the RMS using Spearman's rank correlation (SCC) at three significance thresholds: including residues for which: (i) $P \leq 1$, (ii) $P \leq 0.05$ and (iii) $P \leq 0.05$ after CMT. The SCC was recorded for each of the 18 MSAs, and the mean and standard error (SE) of the means were calculated.

For evaluating threshold-independent performance, the AUC measure was applied. The AUC was calculated against two sets of targets: RMS_{bin} and the enriched set of targets $(RMS+IAS)_{mut}$. The mean AUC and SE were calculated for each set of targets.

Finally, the sensitivity, specificity and MCC were calculated at the same thresholds as the SCC against the enriched set of targets $(RMS+IAS)_{mut}$. The sensitivity, specificity and MCC were recorded for each of the 18 MSAs, and the means and SEs were calculated.

Comparing SigniSite and SPEER

To compare the performance of *SigniSite* with that of existing methods, we turned to a 2009 benchmark study by Chakrabarti and Panchenko (18) comparing the predictive performance of five SDP prediction methods, on a set of protein families with experimentally annotated SDPs. As SPEER (5,19) in this benchmark was found to be the best performing method, we here limit our analysis to comparing *SigniSite* and SPEER by applying both methods to their respective GPDs.

SPEER outputs positional predictions, whereas *SigniSite* assigns a Z-score for each residue type at each position. To cast the *SigniSite* Z-scores into one score per positions, the maximum of the absolute Z-scores was chosen.

SigniSite assigns a prediction value to all positions regardless of residue composition, whereas SPEER by default will skip any fully conserved and positions with >20% gaps. To get prediction values for all positions, we assign a value of '−100' to positions not predicted by SPEER (this value is lower than any score predicted by SPEER).

SPEER requires each sequence in an MSA to be subgroup-annotated before analysis. To accommodate this requirement, each HIV MSA was split into two subgroups, by sorting the sequences in the MSA descending on their associated real values and then splitting the sequences into subgroup '1' or '2' on the median of the sorted values.

To perform the rank analysis *SigniSite* requires that each sequence in the MSA has an associated real number. Of the 20 SDP MSAs, 13 contain only subgroups '1' and '2'. We chose to use these 13 MSAs for the benchmark, using '1' or '2' as '*SigniSite* real number values'.

This way the following two comparisons were made: *SigniSite* versus SPEER on the HIV protease data set and *SigniSite* versus SPEER in the SDP data set. The AUC measure was used to quantify the performance of each method on each benchmark data set.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary description of the *SigniSite* Method,

Supplementary descriptions of the benchmark data sets, Supplementary section on the impact of chosen seed for random number generation, Supplementary description of the benchmarks strategy, Supplementary Tables of HIV-1 PIs and abbreviations.

ACKNOWLEDGEMENTS

The authors thank Martin Blythe for coming up with the name *SigniSite*.

FUNDING

National Institutes of Health [HHSN272201200010C]; EU FP7 PepChipOmics: The European Union 7th Framework Program FP7/2007-2013 [222773]; The Center for Genomic Epidemiology (www.genomicepidemiology.org) grant 09-067103/DSF from the Danish Council for Strategic Research; The University of Copenhagen - Program of Excellence. Funding for open access charge: Technical University of Denmark - PhD programme.

Conflict of interest statement. None declared.

REFERENCES

1. Shcherbo,D., Shemiakina,I.I., Ryabova,A.V., Luker,K.E., Schmidt,B.T., Souslova,E.A., Gorodnicheva,T.V., Strukova,L., Shidlovskiy,K.M., Britanova,O.V. *et al*. (2010) Near-infrared fluorescent proteins. *Nat. Methods*, **7**, 827–829.
2. Gnidehou,S., Jessen,L., Gangnard,S., Ermont,C., Triqui,C., Quiviger,M., Guitard,J., Lund,O., Deloron,P. and Ndam,N.T. (2010) Insight into antigenic diversity of VAR2CSA-DBL5ϵ Domain from multiple *Plasmodium falciparum* placental isolates. *PLoS One*, **5**, e13105.
3. Brandt,B.W., Feenstra,K.A. and Heringa,J. (2010) Multi-Harmony: detecting functional specificity from sequence alignment. *Nucleic Acids Res.*, **38**, 35–40.
4. Capra,J.A. and Singh,M. (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, **24**, 1473–1480.
5. Chakrabarti,S., Bryant,S.H. and Panchenko,A.R. (2007) Functional specificity lies within the properties and evolutionary changes of amino acids. *J. Mol. Biol.*, **373**, 801–810.
6. Kalinina,O.V., Novichkov,P.S., Mironov,A.A., Gelfand,M.S. and Rakhmaninova,A.B. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res.*, **32**, W424–W428.
7. Pei,J., Cai,W., Kinch,L.N. and Grishin,N.V. (2006) Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, **22**, 164–171.
8. Ye,K., Feenstra,K.A., Heringa,J., Ijzerman,A.P. and Marchiori,E. (2008) Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. *Bioinformatics*, **24**, 18–25.
9. Buslje,C.M., Teppa,E., Domnico,T.D., Delfino,J.M. and Nielsen,M. (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. *PLoS Comput. Biol.*, **6**, e1000978.
10. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
11. Dudoit,S., Yang,Y.H., Callow,M.J. and Speed,T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.*, **12**, 111–139.
12. Katoh,K., Misawa,K., Kuma,K.I. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
13. Thomsen,M.C.F. and Nielsen,M. (2012) Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.*, **40**, W281–W287.
14. Lund,O., Nielsen,M., Lundegaard,C., Kesmir,C. and Brunak,S. (2005) *Immunological Bioinformatics*. The MIT Press, Cambridge, MA, London, England.
15. Rhee,S.Y., Gonzales,M.J., Kantor,R., Betts,B.J., Ravela,J. and Shafer,R.W. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.*, **30**, 298–303.
16. Shafer,R.W. (2006) Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.*, **194**, S51–S58.
17. Johnson,V.A., Calvez,V., Gnthard,H.F., Paredes,R., Pillay,D., Shafer,R., Wensing,A.M. and Richman,D.D. (2013) Update of the drug resistance mutations in HIV-1: March 2013. *Top Antivir. Med.*, **21**, 6–14.
18. Chakrabarti,S. and Panchenko,A.R. (2009) Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC Bioinformatics*, **373**, 801–810.
19. Chakraborty,A., Mandloi,S., Lanczycki,C.J., Panchenko,A.R. and Chakrabarti,S. (2012) SPEER-SERVER: a web server for prediction of protein specificity determining sites. *Nucleic Acids Res.*, **40**, W242–W248.
20. Healy,D.G., Falchi,M., O'Sullivan,S.S., Bonifati,V., Durr,A., Bressman,S., Brice,A., Aasly,J., Zabetian,C.P., Goldwurm,S. *et al*. (2008) Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. *Lancet Neurol.*, **7**, 583–590.
21. Dendrou,C.A., Plagnol,V., Fung,E., Yang,J.H., Downes,K., Cooper,J.D., Nutland,S., Coleman,G., Himsworth,M., Hardy,M. *et al*. (2009) Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat. Genet.*, **41**, 1011–1015.
22. Armitage,P., Berry,G. and Matthews,J.N.S. (2002) *Statistical Methods in Medical Research*. Blackwell Publishing Company, Malden, MA, USA.

Part III

Paper II: Insight into Antigenic Diversity of VAR2CSA-DBL5ε Domain from Multiple Plasmodium falciparum Placental Isolates

[On the future of the Bill & Melinda Gates Foundation] "We are committed to the diseases that affects the poorest, malaria, tuberculosis, HIV, all the childhood conditions and until we treat the health of that poor child as being as important as the health of a rich child we will still have work to do"

Bill Gates

3

Insight into Antigenic Diversity of VAR2CSA-DBL5ε Domain from Multiple Plasmodium falciparum Placental Isolates

3.1 Introduction

The work in this paper is based on the fact that immunity towards placental malaria is gradually acquired as a function of parity. The aim of the sequence analysis was to identify parity dependent sequence motifs. I.e. if a given sequence motif is found exclusively in primigravidae women, this means that the motif most likely is immunogenic and an antibody response has been raised against this particular motif. The result of the raised antibody response, is that upon the second pregnancy, infecting parasites expressing a VAR2CSA-DBL5ε variant

containing this particular motif is prevented from cyto-adherence and thus will undergo splenal destruction, the result of which is that the sequence motif will be found in multigravidae women. The extrapolation of the sequence motif identification naturally is identifying vaccine candidates.

3.1.1 Materials

The data set consisted of 70 VAR2CSA-DBL5ε sequences. For each of these sequences the following phenotypes had been recorded: *Maternal age, peripheral parasite conc. [/µl], placental parasite conc. [/µl], maternal blood type, the parity, child gender, child birth weight [g], mean CSA and mean CSPG binding densities [/mm^2]*. It should be noted that not all of the phenotypes had been recorded for each sequence. *SigniSite* was used for parity dependant phenotypic sequence motif identification. Furthermore each of the numerical phenotypes were analysed for parity dependancies.

3.1.2 Methods

Where pairwise complete observations existed, the numerical recorded phenotypes, were correlated with parity and Pearson's correlation coefficient (PCC) [**?**] and Spearman's correlation coefficient (SCC) [**?**] were recorded. Furthermore, the data was split into two groups, such that group 1 contained all phenotypic data from primigravidae women and group 2 all phenotypic data from multigravidae women. Upon this split, the phenotypic data in group 1 and group 2 was compared using 'Welch Two Sample t-test' [4] and 'Wilcoxon rank sum test with continuity correction' [4], the resulting *p-values* were recorded.

3.1.3 Results

Table 3.1.1 summarise results from the parity-correlation of phenotypic values. Only maternal age is found to correlate with parity.

Parity vs.	Age	$Load_{peri}$	$Load_{plac}$	Birth weight	CSA	CSPG
PCC	0.758	-0.389	-0.188	0.135	0.108	0.0802
SCC	0.671	-0.402	-0.248	0.218	0.118	0.0811

Table 3.1.1: VAR2CSA-dbl5ε phenotypic correlations.

Table 3.1.2 summarise the results of testing for phenotypic differences between primigravidae and multigravidae women.

Par_0 vs. $Par_{>0}$	Age	$Load_{peri}$	$Load_{plac}$	Birth weight	CSA	CSPG
t-test	0.000403	0.271	0.0781	0.0168	0.931	0.902
Rank-test	0.00275	0.252	0.0474	0.0592	0.679	0.566

Table 3.1.2: VAR2CSA-dbl5ε primi- vs. multigravidae phenotypic comparisons. *p-values* calculated using 'Welch Two Sample t-test' and 'Wilcoxon rank sum test with continuity correction' are stated. Significant *p-values* at a level of significance of $\alpha = 0.05$ are highlighted in red.

3.1.4 Discussion

Significant phenotypic differences are identified between primigravidae and multigravidae women, based on maternal age, parasite load in the placenta and birth weight of the child. All of these findings are in line with the gradual acquisition of immunity towards PM. The parasite load is reduced, with increasing immunity and as a result hereof the fetus receives more nutrition and is thus able to obtain a higher birth weight.

One major challenge in the analysis of this set of VAR2CSA-DBL5ε data, was the lack of data. A total of 70 VAR2CSA-DBL5ε sequences lay the foundation for the anaylsis, however as several isolates were lacking phenotypic annotation. The actual number of sequences included in each part of the analysis was as low as 33. The consequence of this naturally is that it is difficult to make extrapolations and conclusions based on a data set of this size.

3.2 Paper II

The following paper was published in PLOS ONE in October 2010 as a collaboration between 'Département Santé. Institut de Recherche pour le Développement (IRD). Faculté de pharmacie, Université Paris René Descartes - Paris 5, Paris, France.' and the Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU). The IRD collected samples and obtained phenotypic profiles and CBS/DTU subsequently performed the *in silico* sequence analysis.

OPEN ACCESS Freely available online

PLoS one

Insight into Antigenic Diversity of VAR2CSA-DBL5ε Domain from Multiple *Plasmodium falciparum* Placental Isolates

Sédami Gnidehou[1,2]*, Leon Jessen[3], Stéphane Gangnard[4], Caroline Ermont[1,2], Choukri Triqui[1,2], Mickael Quiviger[1,2], Juliette Guitard[1,2], Ole Lund[3], Philippe Deloron[1,2], Nicaise Tuikue Ndam[1,5]*

1 Institut de Recherche pour le Développement, IRD UMR 216, Mère et Enfant Face aux Infections Tropicales, Paris, France, **2** Université Paris Descartes, Paris, France, **3** Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark, **4** Unité d'Immunologie Structurale, Institut Pasteur, CNRS URA2185, Paris, France, **5** Institut des Sciences Biomédicale et Appliquées, Cotonou, Benin

Abstract

Background: Protection against pregnancy associated malaria (PAM) is associated with high levels of anti-VAR2CSA antibodies. This protection is obtained by the parity dependent acquisition of anti-VAR2CSA antibodies. Distinct parity-associated molecular signatures have been identified in VAR2CSA domains. These two observations combined point to the importance of identifying VAR2CSA sequence variation, which facilitate parasitic evasion or subversion of host immune response. Highly conserved domains of VAR2CSA such as DBL5ε are likely to contain conserved epitopes, and therefore do constitute attractive targets for vaccine development.

Methodology/Principal Findings: VAR2CSA DBL5ε-domain sequences obtained from cDNA of 40 placental isolates were analysed by a combination of experimental and *in silico* methods. Competition ELISA assays on two DBL5ε variants, using plasma samples from women from two different areas and specific mice hyperimmune plasma, indicated that DBL5ε possess conserved and cross-reactive B cell epitopes. Peptide ELISA identified conserved areas that are recognised by naturally acquired antibodies. Specific antibodies against these peptides labelled the native proteins on the surface of placental parasites. Despite high DBL5ε sequence homology among parasite isolates, sequence analyses identified motifs in DBL5ε that discriminate parasites according to donor's parity. Moreover, recombinant proteins of two VAR2CSA DBL5ε variants displayed diverse recognition patterns by plasma from malaria-exposed women, and diverse proteoglycan binding abilities.

Conclusions/Significance: This study provides insights into conserved and exposed B cell epitopes in DBL5ε that might be a focus for cross reactivity. The importance of sequence variation in VAR2CSA as a critical challenge for vaccine development is highlighted. VAR2CSA conformation seems to be essential to its functionality. Therefore, identification of sequence variation sites in distinct locations within VAR2CSA, affecting antigenicity and/or binding properties, is critical to the effort of developing an efficient VAR2CSA-based vaccine. Motifs associated with parasite segregation according to parity constitute one such site.

Citation: Gnidehou S, Jessen L, Gangnard S, Ermont C, Triqui C, et al. (2010) Insight into Antigenic Diversity of VAR2CSA-DBL5ε Domain from Multiple *Plasmodium falciparum* Placental Isolates. PLoS ONE 5(10): e13105. doi:10.1371/journal.pone.0013105

Editor: David Joseph Diemert, The George Washington University Medical Center, United States of America

Received March 3, 2010; **Accepted** July 29, 2010; **Published** October 1, 2010

Funding: This work was supported by grants from the French National Agency of Research (grant # ANR-05-MIME-009-01), the European Commission, FP7 work program (Grant # 200889), and the Copenhagen University Programme of Excellence. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gcarine@yahoo.com (SG); nicaise.ndam@ird.fr (NTN)

Introduction

Women suffering from pregnancy-associated malaria (PAM) develop antibodies that protect them and their offspring during subsequent pregnancies [1]. Protection against PAM is rapidly acquired as from the second pregnancy, and is associated with increasing plasma levels of PAM-specific anti-Variant Surface Antigen (VSA) antibodies. PAM parasites from distinct geographic areas specifically bind Chondroitin-Sulfate A (CSA) [2,3,4], and the immune response in pregnant women living in malaria endemic areas is highly directed against *var2csa* encoded PfEMP1 (*Plasmodium falciparum* erythrocyte membrane protein) protein [5,6,7]. Protective antibodies in PAM immunity are thought to recognize a relatively conserved antigen that mediates parasite binding to placental CSA, as plasma and parasites from pregnant women of different malaria endemic areas cross-react [5], [8]. Antibodies against VAR2CSA are sex-specific and parity-dependent, and high levels of such antibodies are associated with reduced consequences of PAM, making VAR2CSA a promising target for vaccine development [6,7].

The VAR2CSA protein is a large antigenic molecule (350 kDa), exposed to host antibodies on the surface of erythrocytes [9,10]. It has been shown that disruption of *var2csa* results in the loss of CSA adhesion ability of infected erythrocytes (IE) [11]. The VAR2CSA protein is structurally composed of six Duffy Binding-Like (DBL) domains. Several of these domains, including DBL5ε, have, to some extent, displayed affinity for CSA *in vitro* [12,13,14,15] [16]. Antibodies raised against CSA-binding VAR2CSA domains have so far not been able to exhibit strong adhesion-inhibitory

capabilities. However, antibodies raised against the recombinant DBL5ε domain amplified from a placental parasite, have been shown to bind native VAR2CSA expressed on the surface of *P. falciparum* IEs from placental isolates [16].

Var2csa is a polymorphic gene [17], and intra strain variability represents a great challenge for vaccine development. In a previous study, using genomic DNA from *P. falciparum* parasites from Senegalese women, the DBL5ε domain was found to be highly conserved among parasite isolates [18]. Mapping on a structural model revealed the localization of the DBL5ε identified polymorphic and some conserved regions in the exposed loops and helices [8,18].

Although most VAR2CSA DBL domains contain conserved and polymorphic domain regions that can be targeted by surface reactive antibodies [8], conserved regions are most prominent in DBL3X, DBL4X and DBL5ε. This may explain why antibodies raised against DBL3X and DBL5ε recombinant proteins exhibited most cross-reactivity with heterologous parasites compared to antibodies raised against the other domains [19]. Interestingly, these antibodies (raised against a single variant of DBL3X or DBL5ε) cross-reacted with placental parasite isolates from Tanzania [20]. Moreover, human monoclonal antibodies produced by immortalized B cells from malaria-exposed pregnant women predominantly recognized DBL3X and DBL5ε [21], suggesting the natural acquisition of a specific immune memory to these VAR2CSA domains.

Together, these observations highlight that DBL5ε may represent an interesting target for vaccine development. Understanding the molecular basis controlling the broad and/or differential antibody recognition of this VAR2CSA domain may help define essential structural features of a potential interest in vaccine perspectives. The two main objectives of this study were: (i) To analyse the consequence of sequence variation in the VAR2CSA DBL5ε domain using the transcripts from a large panel of fresh placental parasite isolates and, (ii) to express and to characterize selected VAR2CSA DBL5ε variants from two parasite isolates. Novel conserved linear epitopes which are recognised by naturally acquired antibodies were found in the conserved regions of the DBL5ε domain and significant motifs were identified in the variable regions.

Results

Identification of significant sites in VAR2CSA DBL5ε sequences

Figure 1 shows a multiple alignment of 70 VAR2CSA DBL5ε sequences (All sequence data are available at GenBank (http://www.ncbi.nlm.nih.gov/Genbank) under the accession numbers HM751723–HM751795) using cDNA from 40 placental parasites isolated at delivery from 39 Senegalese women [2,15,22] and one

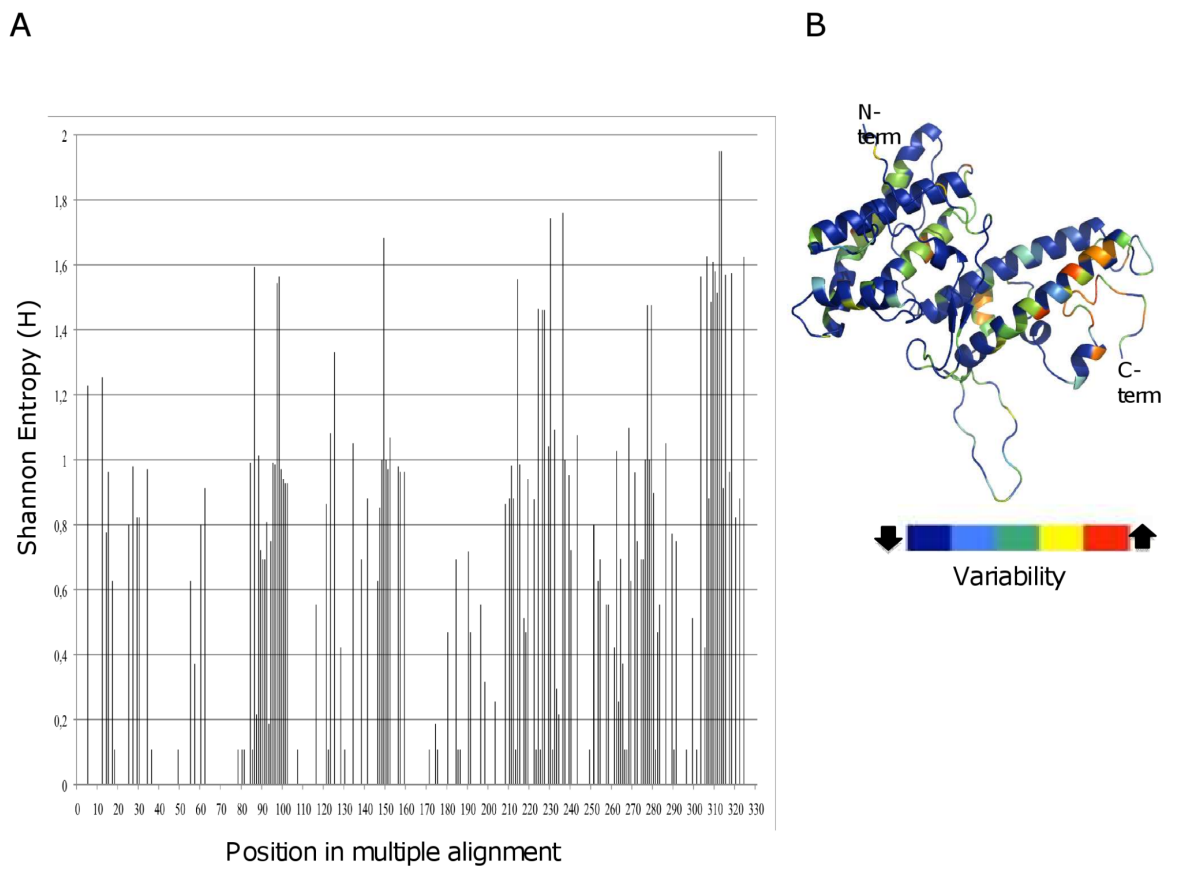


Figure 1. High conservation of DBL5ε-VAR2CSA sequences. (A) Plot of DBL5ε Shannon entropy (H): $H = 0$: Complete conservation, only one residue present at the given position. $0 < H \leq 1$: Considered highly conserved. $1 < H \leq 2$: Considered conserved. $2 < H \leq 4.3$ considered variable. (B) Three-dimensional model of DBL5ε showing the sequence variability. Heat-map colouring is dark blue (conserved) to red (variable).
doi:10.1371/journal.pone.0013105.g001

Tanzanian woman [20]. The var2csa region corresponding to DBL5ε plus Id5 (the non-DBL Interdomain sequence located between DBL5ε and DBL6ε) was cloned and sequenced. A total of 70 VAR2CSA DBL5ε sequences were obtained from these 40 placental parasites. The multiple alignment of DBL5ε sequences using the calculated Shannon entropy values show that the sequences consist of constant and variable blocks (Figure S1, Figure 1A). Conservation of 85% was obtained with DBL5ε and 80% when we considered DBL5ε plus Id5. The variability mapping on the DBL5ε structural model revealed that conserved and variable areas were located in loops and protruding helices (Figure 1B). In a previous study, it was found that VAR2CSA DBL3X sequence motifs can be linked to the parity of the infected women [15]. In order to assess such sequence variation behaviour in another highly immunogenic and conserved VAR2CSA domain, all DBL5ε sequences generated from cDNA of PAM isolates were analysed using SigniSite [23]. Analysis revealed that certain amino acids of VAR2CSA DBL5ε+Id5 sequences appear to be of particular interest. In the multiple alignment of all DBL5ε sequences, significantly distributed residues were identified at positions 277, 279, 303 (Fig 2A and 2B). High CSPG (Chondroitin Sulphate Proteoglycan) binding density is correlated with amino

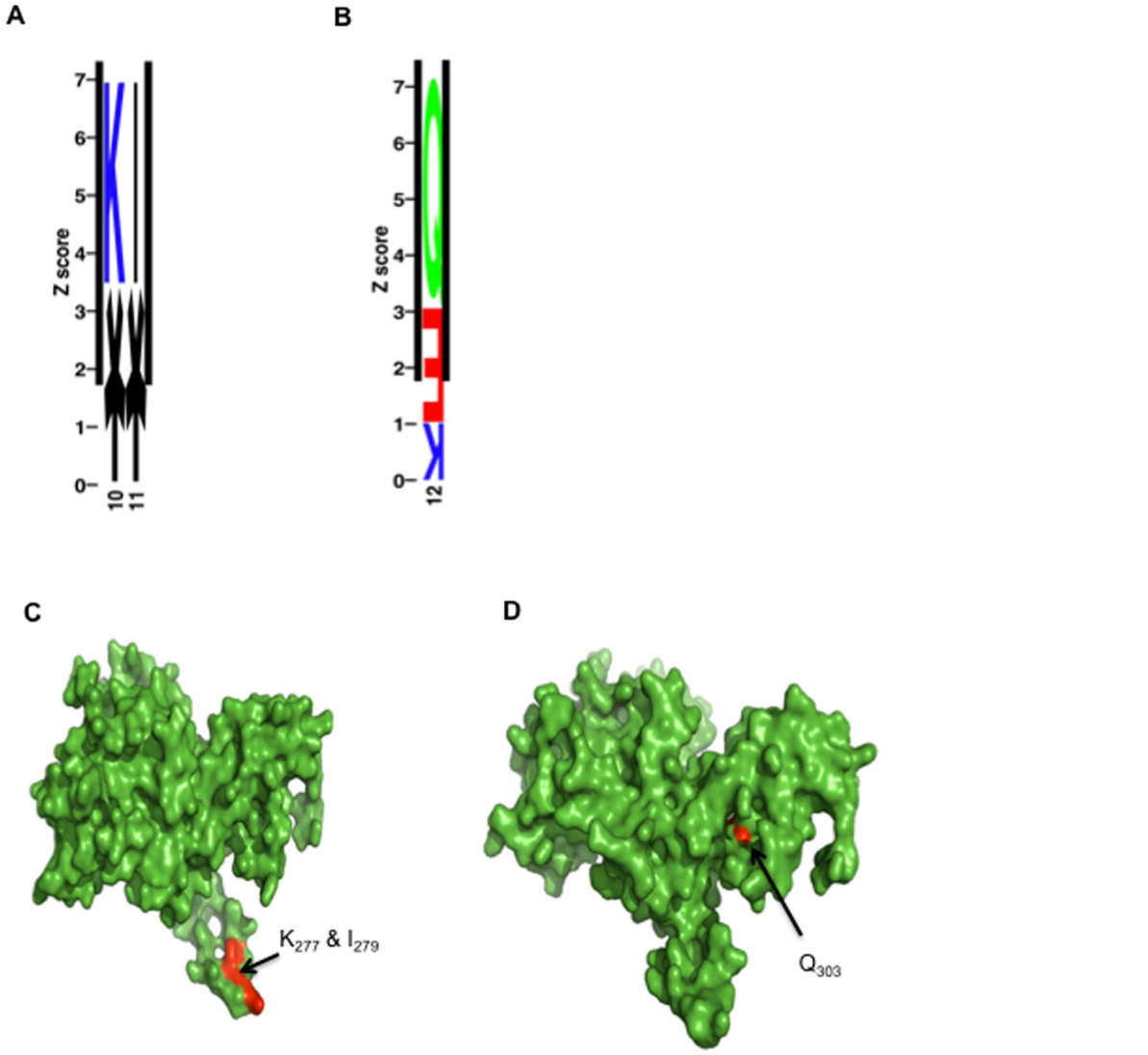


Figure 2. VAR2CSA DBL5ε patterns distribution. (A, B): Sequence logo showing the identified significantly distributed residues I, K and Q The sequence logo displays the residues present at each position, where at least one residue was identified as being significantly distributed with respect to associated numerical parameter. Each letter denotes a given residue and the height corresponds to increasing z-score. The residues are coloured according to: Acidic [ED]: red, Basic [RKH]: blue, Neutral [GNQSTY] = green, Hydrophobic [ACFILMPVW] = black. Numbers below each column denotes corresponding position in the multiple alignment. Letters positioned correctly are associated with high values and upside down letters with low. An asterisk denotes a deletion. It should be noted that in the sequence logos other residues appears (*, E, K), these are however not identified as significantly distributed (i.e. $p>0.05$). DBL5ε models showing the position of the identified significant residues (red), T_{277}, I_{279} (C) and Q_{303} (D).
doi:10.1371/journal.pone.0013105.g002

acid Q_{303} ($p = 0.017$). Homology modelling of DBL5ε-3d7 furthermore revealed that identified residues that were significantly different among groups were surface-exposed (Figure 2C, Figure 2D).

From visual inspection of the regions around the amino acid residues found by SigniSite analysis in the multiple alignment of DBL5ε, distinct motifs were identified when comparing sequences from primigravidae and multigravidae. Motifs VFNNA, gap, TFKNI were identified in the area spanning amino acids 275 to 279 and EDTKQ, EYTGN and QYTGN were defined between the amino acid 303 and 313 (this area is located at the end of DBL5ε and in Id5). These patterns have a differential distribution according to parity. Indeed, gap, EDTKQ and EYTGN motifs were predominantly found among samples from primigravidae ($p = 0.02$, Fisher's exact test) whereas TFKNI and QYTGN were mainly or exclusively found in multigravidae ($p = 0.013$). These patterns clearly discriminate parasites infecting multigravidae and primigravidae women. At the level of sequence types obtained from each sample, DBL5ε sequences expressing gap, EDTKQ and EYTGN signatures were found mostly in primigravidae ($p = 0.036$) while those expressing TFKNI ($p = 0.0019$) and/or QYTGN preferentially infect ($p = 0.038$) multigravidae (Table 1). Interestingly the TFKNI motif was also associated with high maternal age and low placental parasite density (data not shown). The VFNNA motif was found in primigravidae as well as multigravidae without significant bias in its distribution. From the mapping of TFKNI and deletion motifs on the DBL5ε structural model from multigravidae CYK008 sequence and primigravidae CYK040 respectively, it can be hypothesised that TFKNI insertion can cause a conformational change of the domain structure (Figure 3).

Expression of distinct variants of recombinant VAR2CSA DBL5ε from placental parasites

Two VAR2CSA DBL5ε variants (CYK39 and CYK49) were produced in *Rosetta gami* DE3 strains. Both variants were chosen for analysis as *P. falciparum* IEs corresponding to isolate CYK39 have been described as high CSPG binders and parasites from CYK49 as low binders [2]. The *Rosetta gami* bacteria strain allows the formation of disulfide bonds that could favour production of biologically active proteins. Protein production was induced with 0.1 and 1 mM IPTG. The soluble protein produced was affinity-purified, subjected to gel filtration, and the purity was checked by SDS-PAGE (Figure 4A) and Western blotting. An average of 5 mg of pure protein was obtained after the different purification steps. Western blot analysis showed that total IgG purified from a plasma pool of malaria exposed multigravidae labelled a single dominant band of 37 kDa in 1 mM IPTG induced bacterial extract and in purified DBL5ε (Figure 4B). The same product (37 kDa) was identified by specific IgG generated in mice by DNA vaccination with DBL5ε_CYK39 (Figure 4C) and DBL5ε_ CYK49 IgG (Figure 4D), as well as with anti-histidine tag monoclonal antibodies (Figure 4E). Bands of expected size were observed neither in the untransformed nor uninduced bacterial extracts (Figure 4).

In vitro binding ability of placental parasite recombinant DBL5ε VAR2CSA variants to CSPG

The CSPG binding capacity of the two DBL5ε variants was estimated by ELISA. Both variants showed a relatively higher binding ability to CSPG compared to NTS-DBL1α domain of VARO (Figure 5A). This interaction was concentration-dependent. In this model, the NTS-DBL1α domain of VARO also produced in *Rosetta gami* displayed weak binding ability to CSPG. To determine whether this interaction was CSPG-specific, we tested the ability of soluble CSPG (decorin) or CSA (bovine trachea CSA) to compete for protein binding on a CSPG precoated plate. As shown in Figure 5B, soluble CSPG like soluble CSA (data not shown) indeed competed for binding observed on CSPG. Sequence comparison of both DBL5ε variants expressed showed that they were highly similar but contained 31 different residues. Moreover, positively charged amino acids appeared to be differentially expressed in both variants (Figure 5C). As position 303 seemed to be associated with binding density, the sequences were analysed for mean CSA and CSPG binding densities and the difference associated with the occurrence of the Q, E and K residues. Indeed, high CSA or CSPG binding affinity was mainly associated with residue Q_{303} ($p = 0.005$) whereas low CSA or CSPG binding affinity was associated with E/K_{303} (Table 2). Interestingly as shown in figure 5C, the equivalent residue for CYK39 and CYK49 sequences was in fact $Q_{296\ and}$ K_{296} respectively. The mapping of Q_{303} on structural model indicates that this residue seems to be surface exposed, but located in the bottom of what could be a binding pocket (Figure 2D).

Antibodies against DBL5ε domain of VAR2CSA increase in a parity-dependent manner

Recombinant DBL5ε variants (CYK39 and CYK49) were used to assess the plasma levels of anti-VAR2CSA IgG. Independent of which variant was tested, antibodies with specificity for *Rosetta gami*-produced DBL5ε VAR2CSA were seen only in plasma from *P. falciparum*-exposed pregnant women living either in Benin (Ben) or in

Table 1. Signatures in DBL5ε domain of VAR2CSA expressed by placental parasites.

Category	Parity	VAR2CSA DBL5ε motifs					
		VFNNA	Gap	TFKNI	EDTKQ	EYTGN	QYTGN
Samples	Primigravidae (n = 16)	6	12[a]	1	7[a]	6[a]	0
	Multigravidae (n = 24)	7	9	11[a]	2	2	5
Sequences	Primigravidae (n = 33)	11	20[a]	1	7[a]	8[a]	0
	Multigravidae (n = 37)	11	14	12	2	2	5[a]

Gap, EDTKQ and EYTGN motifs are mainly found in samples from primigravidae compared to those from multigravidae ($p = 0.02$) whereas TFKNI and QYTGN are mostly or exclusively found in multigravidae ($p = 0.013$). At the level of sequence types obtained from each sample, the EDTKQ, EYTGN and gap signatures are more frequent among DBL5ε sequences from primigravidae compared to those originating from multigravidae ($p = 0.036$). Similarly, the TFKNI and QYTGN motifs are mostly or exclusively found in sequences from multigravidae ($p = 0.0019$ and $p = 0.038$ respectively).

[a]$p < .05$, Fisher's exact test.

[b]$p < .001$, Fisher's exact test.

doi:10.1371/journal.pone.0013105.t001

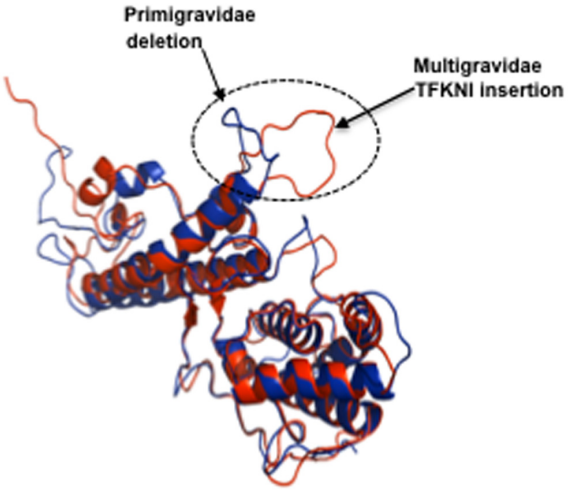


Figure 3. Mapping of VAR2CSA-DBL5ε signatures. Based on the identified region of interest and predominant motifs, two representative sequences were selected for homology modelling primigravidae CYK040 (deletion) and multigravidae CYK008 (TFKNI). Blue is CYK040 primigravidae sequence, red is CYK008 multigravidae sequence and dotted circle is deletion/TFKNI motif. The figure illustrates how the conformation of the region depends on the presence or absence of the TFKNI-motif. Using homology modelling, the motif is identified as being surface exposed and may thus alter the immunogenicity of the region.
doi:10.1371/journal.pone.0013105.g003

Senegal (Sen) (Figure 6A). In contrast, plasma levels of antibodies against the recombinant DBL5ε were insignificant in both French unexposed men (M) and pregnant women (Fra) (Figure 6A). Detailed analysis of *P. falciparum*-exposed pregnant women indicated that for each antigen tested, Senegal and Benin multigravidae (M) had significantly higher levels of DBL5ε antibodies than primigravidae (P); (CYK39: both $p<0.0001$; CYK49: $p=0.019$ for Senegalese and $p<0.0001$ for Beninese; Figure 6B), however contrary to Senegalese primigravidae, most Beninese primigravidae presented with high DBL5ε VAR2CSA antibody levels (Figure 6B). A fine analysis of the plasma reactivity of the women demonstrates that antibodies against DBL5ε increased with parity (Figure 6C). We compared plasma levels of VAR2CSA specific IgG using both DBL5ε recombinant proteins. Cut-off values were set to mean + 2SD (plus two standard deviations) of reading obtained with the negative control plasma samples. The percentage of antibody reactivity considered to be positive was 80% for DBL5ε_CYK39 and 60% for DBL5ε_CYK49. Despite a homology of 80%, there is a significant difference of reactivity between both variants (Chi^2 test $p=0.005$). A comparative study of the reactivity of each plasma with respect to each of the variants shows that the response to both variants was strongly correlated (Pearson's test $r=0.8$, $p<.0001$; Figure 6D), confirming that the VAR2CSA DBL5ε domain contained conserved epitopes.

Evidence of conserved cross-reactive epitopes in DBL5ε VAR2CSA

Recombinant DBL5ε variants were used in competition ELISAs to demonstrate that DBL5ε domain of VAR2CSA contains cross-

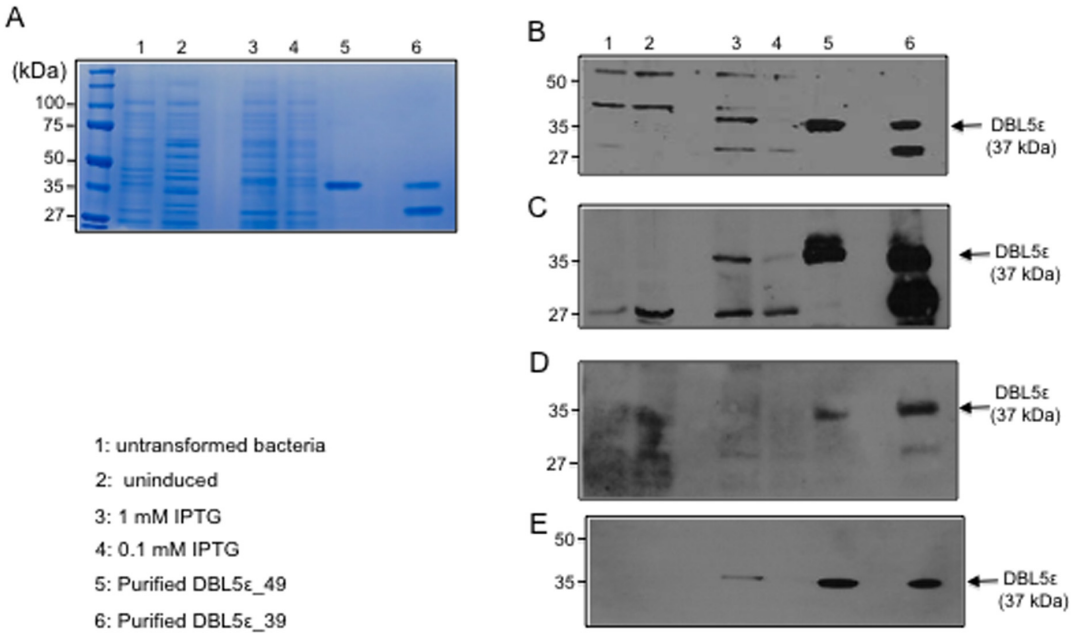


Figure 4. Bacterial recombinant DBL5ε domain of VAR2CSA expression. Lysates of untransformed (lane 1) bacteria, DBL5ε_CYK49 [uninduced (lane 2), induced 1mM IPTG (lane 3), induced 0.1 mM IPTG (lane 4)], DBL5ε_CYK49 (lane 5) and DBL5ε_CYK39 (lane 6) after two purification steps were subjected to SDS/PAGE and either stained with Coomassie blue (A) or immunoblotted with either purified IgG multigravidae plasma (B), antisera from mice vaccinated with DBL5ε_CYK39 (C), antisera from DBL5ε_CYK49 vaccinated mice (D) or monoclonal anti-histidine antibodies (E). 30 μg of bacteria-expressed-extract proteins and 2 μg of purified domains were used for analysis. Immune complexes were detected with appropriate horseradish peroxidase coupled antibodies.
doi:10.1371/journal.pone.0013105.g004

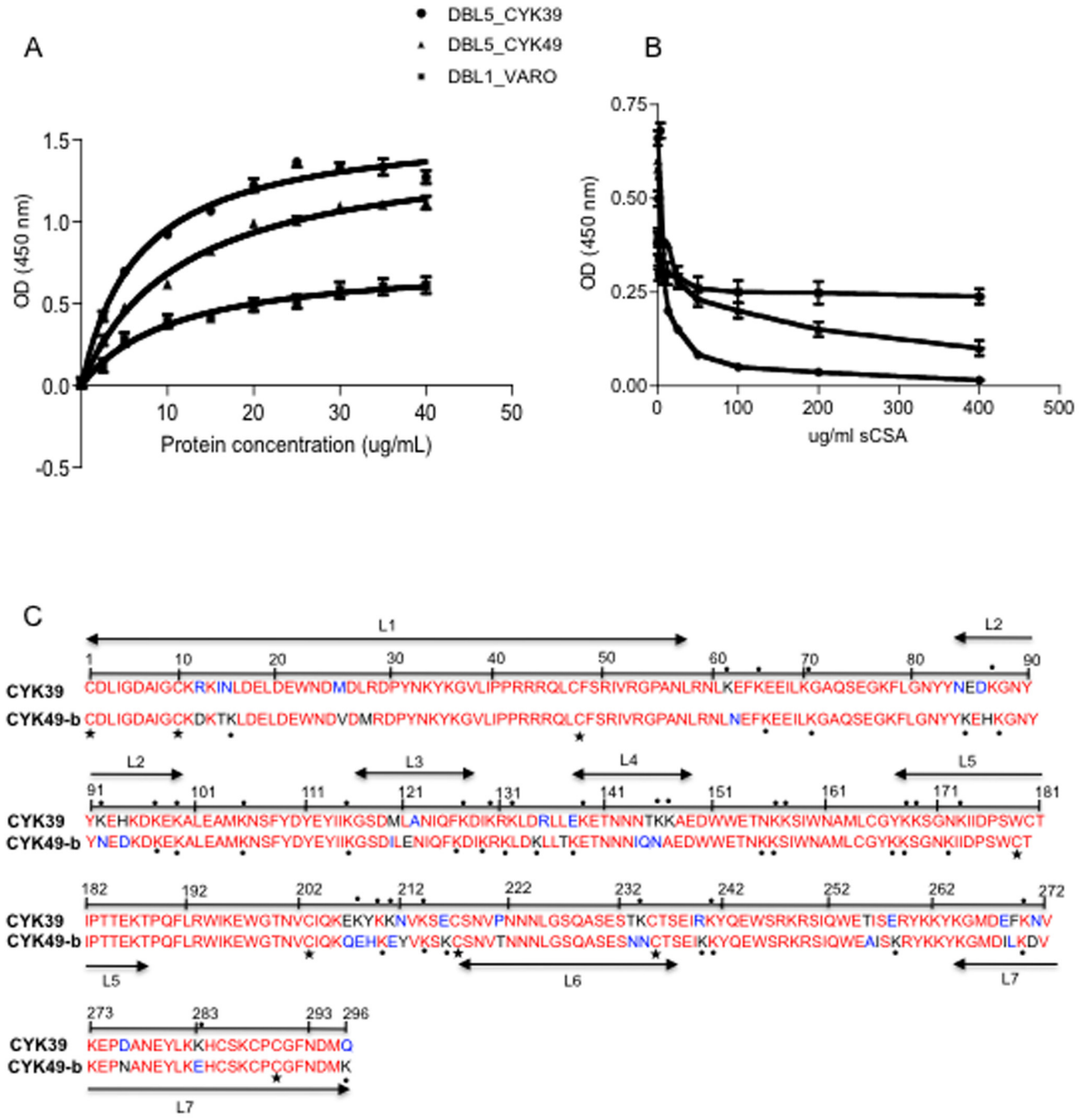


Figure 5. CSPG binding of the DBL5ε domain of the VAR2CSA from parasite isolates. (A): Increasing concentrations of protein were added to wells coated with 5 μg/ml of CSPG. CSPG-binding of the DBL5ε_CYK39 (circle), DBL5ε_CYK49 (triangle) and the non CSA-binding VARO NTS-DBL1α domain used as control (square). Results are the means of three binding assays and the error bars indicate the standard deviation. (B) Inhibition assay. Recombinant DBL5ε variants (5 μg/ml) were pre-mixed with increasing amounts of soluble CSA 0.25–400 μg/ml, and binding to CSPG-coated plates was determined. Results are the means of three inhibition binding assays and error bars indicate the standard deviation. (C): Sequence comparison of VAR2CSA DBL5ε domains from CYK39 and CYK49. Asterisks and circles indicate respectively Cystein residues and Lysine. Conserved amino acids are shown in red and polymorphic residues in black. The 7 loops (L1–L7) identified according to Andersen P *et al.* [8] are underlined.
doi:10.1371/journal.pone.0013105.g005

reactive epitopes. While one variant of the two expressed VAR2CSA DBL5ε was used for coating, the other one was used as soluble competitor. The antibody reactivity of either a high-titered VAR2CSA plasma pool from Beninese or Senegalese women, or antisera to DBL5ε_CYK39 and DBL5ε_CYK49 generated in mice by DNA vaccination, or plasma pool from unexposed French pregnant women was compared with or without pre-incubation with increasing concentrations of the competing VAR2CSA DBL5ε variant. As negative control, all plasma were incubated with VARO NTS-DBL1α domain. Figure 7 shows that DBL5ε from placental parasites contains conserved epitopes. Indeed, whichever the DBL5ε variant tested,

Table 2. VAR2CSA-DBL5ε residues Q_{303}, E_{303} and K_{303} distribution in relation to placental parasite CSA/CSPG binding affinity.

	Q_{303}	E_{303}	K_{303}
Isolates			
High binders (n = 20)	10/20[a]	6/20	6/20
Low binders (n = 16)	1/16	7/16	9/16

Those parasites ability to bind CSA or CSPG have previously been described [2].
[a]$p<.01$, t-test.
n corresponds to placental parasite isolates.
doi:10.1371/journal.pone.0013105.t002

the competitor inhibited its antibody recognition in a concentration-dependant manner (Figure 7A). No significant competition was seen with the negative VARO control protein (Figure 7B). Due to the highly conserved nature of VAR2CSA DBL5ε sequence, it was decided to determine whether any of its conserved regions was recognised by naturally acquired antibodies. We synthesised a library of peptides using 3D7 DBL5ε sequence. All peptides were screened in ELISA for reactivity against a plasma pool from Beninese or Senegalese women, French unexposed pregnant women and men. Two peptides P4 and P13 located in highly conserved regions of VAR2CSA displayed significant and specific recognition by plasma of malaria exposed pregnant women compared to control plasma from French unexposed pregnant women and men (Kruskal-Wallis test, $p<0.0001$; Figure 8A). Antibody reactivity of both peptides was higher in multigravidae compared to primigravidae, though not significant (Mann-Whitney U, $p = 0.17$).

Specific antibodies to VAR2CSA DBL5ε conserved peptides mark native VAR2CSA on the surface of infected erythrocyte

Mapping of both peptides P4 and P13 on DBL5ε structural model indicated that both of them are surface-exposed (Figure 8B). Furthermore, specific antibodies against both peptides were affinity-purified from the Senegalese pregnant women plasma pool and allowed to react with PAM parasites collected from pregnant women from Benin. The pregnancy specific antibody recognition of the isolates used was checked prior by FACS with human plasma control pools (data not shown). The results presented on Figure 8C show that the antibodies with specificity to the selected peptides reacted with the native VAR2CSA expressed by PAM parasites on the surface of IE.

Discussion

Pregnant women acquire protective antibodies that cross-react with geographically diverse placental *P. falciparum* isolates, suggesting that surface molecules expressed on infected erythrocytes (IE) by PAM parasites have conserved epitopes and, thus, that a PAM vaccine may be possible to achieve. The search for surface antigens of placental *P. falciparum* parasites is focused on the PfEMP1 family. Most studies in recent years have shown that VAR2CSA is the dominant PfEMP1 associated with parasite binding to the placenta. Due to technological difficulties the exact conformation of the entire VAR2CSA protein remains unknown. Preliminary studies to understand its binding properties focused on its DBLs domains and functional studies have shown that several VAR2CSA DBLs including DBL5ε can individually bind CSA *in vitro*. This approach has become questionable as no efficient anti-adhesion antibodies for IE have been obtained following vaccination with a single domain. Recent studies have nevertheless demonstrated that VAR2CSA DBL5ε domain can induce antibodies with a broad range of reactivity against placental isolates [16,20] and therefore may represent a potential target for PAM vaccine development. This study analysed sequence variation in the DBL5ε domain of the transcribed *var2csa* gene from multiple placental parasite isolates. The aim was to evaluate antigenic diversity and diversifying pressure within this attractive VAR2CSA area. Using cDNA (complementary acid deoxyribonucleic) from 40 placental parasite isolates from a previous study, the region encoding DBL5ε+Id5 of *var2csa* was amplified, cloned and sequenced. Findings from our study population clearly confirmed previous observations that the VAR2CSA DBL5ε is highly conserved [18]. Indeed, an average of 81% amino-acid sequence identity was seen among DBL5ε sequences as reported by Guitard et al. on a different study population [18]. Variations were mainly located in segments of variable length and mapping of DBL5ε regions to 3-D model revealed that variable areas are located in the loops and protruding helices [8].

Two variable regions, one in the DBL5ε and another one in the Id5 sequences appeared to be of particular interest regarding the bias in motif distribution among gravid women. Three significant motifs (gap, VFNNA and TFKNI) were identified in the first region spanning Aa from position 275 to 279. Despite the relatively high variability of the Id5, another area with motif segregation (EDTKQ, EYTGN and QYTGN) was found between Aa 303 and 313. The major observation in these sites is the significant difference between motif occurrence among parasites from primigravidae and multigravidae. Certain motifs are preferentially found in parasites from primigravidae ($gap_{275–279}$, $E_{303}D_{308}T_{309}K_{312}Q_{313}$ and $E_{303}Y_{308}T_{309}G_{312}N_{313}$), whereas others are only found in parasites infecting multigravidae ($TFKNI_{275–279}$ and $QYTGN_{303–313}$). Interestingly, most of the parasites with $QYTGN_{303–313}$ motif also had $TFKNI_{275–279}$. Those expressing either $EDTKQ_{303–313}$ or $EYTGN_{303–313}$ are mostly associated with a $gap_{275–279}$. Such selection pattern was already seen in the DBL3X sequence and plausible explanations can be given, based on several hypotheses: (i) either that parasites infecting primigravidae are the most efficient mediators for binding and therefore have a biological advantage in women with limited immunity against PAM, (ii) or that the parasite variants mostly found in primigravidae are the more common in the area and therefore are more likely to infect exposed primigravidae while multigravid women already have developed specific antibodies during previous pregnancies. The tropism of certain parasite variants for multigravid women suggest that some rarer variants, probably not the most virulent can escape existing immunity to common VAR2CSA variants. These findings have important implications for understanding immunity to PAM in a context where the development of a VAR2CSA-based vaccine is gaining interest. Further analyses in this study also found a significant difference at a site situated in the Id5 according to the ability of IE to bind CSA or CSPG *in vitro*. Isolates with high binding affinities associated with Q_{303} and low CSA/CSPG binders associated with E/K_{303}. This could indicate that conservation of Q_{303} may have conformational importance for maintaining high binding ability by the IE.

The results generated in the present study highlight the fact that fundamental gaps remain in our knowledge and understanding of placental parasites. Protection against PAM is consistent with repeated exposure during pregnancy to previously unknown antigens. Most of multigravidae infected by parasites with the TFKNI or QYTGN motifs have a parity status above 3,

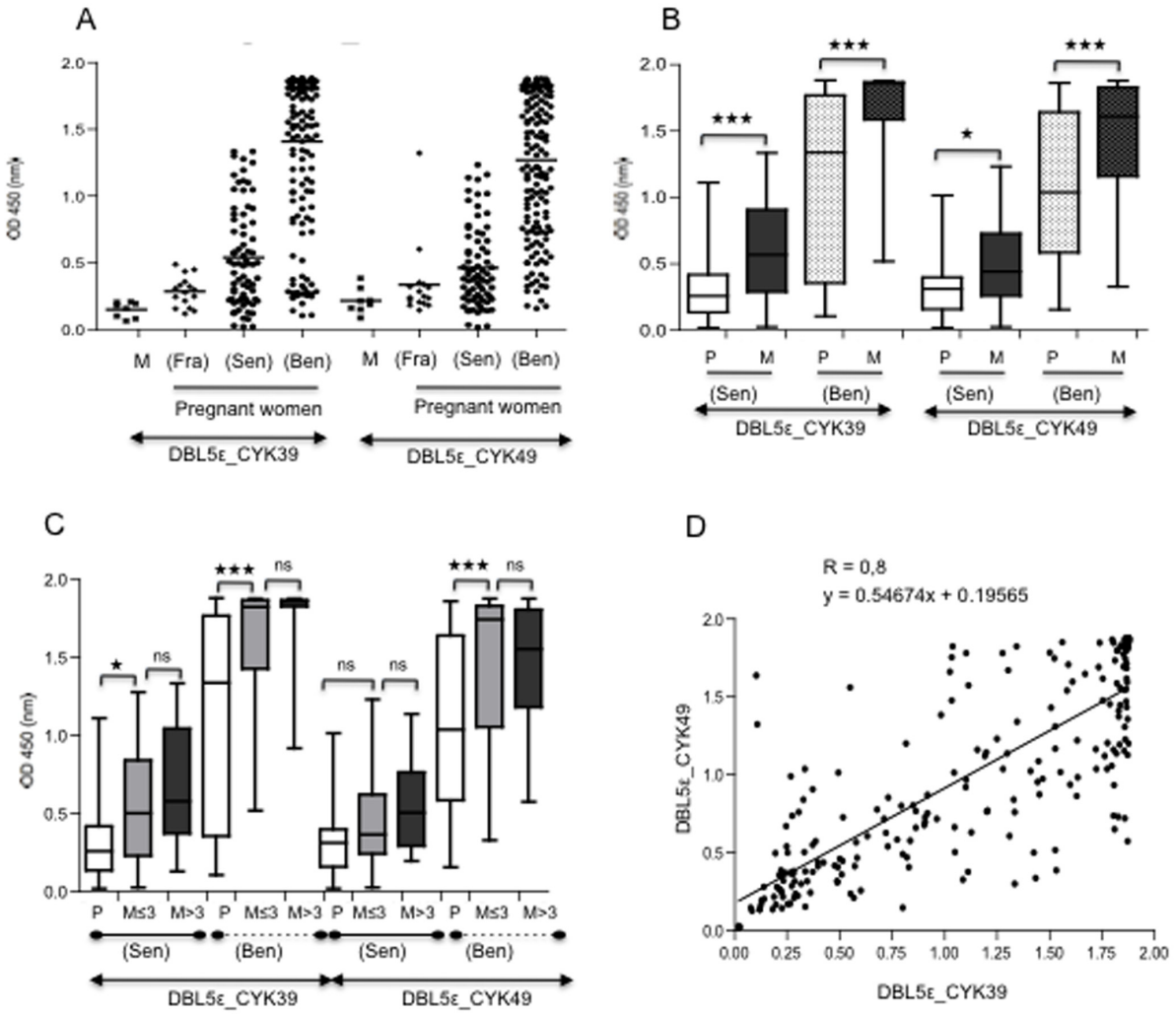


Figure 6. Plasma reactivity against DBL5ε domains of VAR2CSA. (A): Plasma levels of IgG with specificity for DBL5ε domain of VAR2CSA in 8 French unexposed men (M), 16 French unexposed pregnant women (Fra), 75 Senegalese pregnant women (Sen) and 160 Beninese pregnant women (Ben). DBL5ε variants CYK39 and CYK49 were tested. (B): Plasma levels of VAR2CSA DBL5ε domain according to parity. DBL5ε antibodies levels were quantified in the same groups of malaria-exposed pregnant women (Benin and Senegal) as in A. 24 primigravidae (P), 51 multigravidae (M) from Senegal; 80 primigravidae and 80 multigravidae from Benin. (C): Plasma levels of VAR2CSA DBL5ε domain according to parity range. Malaria exposed women used in (B) were separated in three groups; primigravidae (P), women whose parity level is lower or equal to 3 (M≤3) [Beninese women: n = 48, n = 26 for Senegalese women] and those whose parity status is higher than 3 (M>3) [Beninese women: n = 32, n = 25 for Senegalese women]. (D) Correlation between the reactivity to each DBL5ε variant in a given plasma.
doi:10.1371/journal.pone.0013105.g006

suggesting that despite the protection acquired during different pregnancies, women can still be infected by new parasite variants [18]. In the context of developing an optimal VAR2CSA-based vaccine that can protect against placental malaria, it will be particularly useful to overcome the challenges associated to sequence variation in this interesting candidate. The relation underlying the even limited variations described in this study suggests that these can have critical implication in the functionality of the whole molecule including its ability to subvert immunity. Our results clearly indicate that the design of a protective vaccine based on VAR2CSA should not be limited to a single variant. A limited number of variants may be sufficient for broad coverage, provided sites under significant variations are considered.

We have characterized two distinct variants of DBL5ε from our study population. The measure of plasma levels of the antibodies against these two DBL5ε variants showed that the two proteins were broadly recognized by samples from two malaria endemic regions with different *P. falciparum* transmission levels. Both VAR2CSA DBL5ε variants were recognized in a parity-dependent manner although the acquisition of immunity against VAR2CSA differed between the two regions. In areas of intense *P. falciparum* transmission, pregnant women generally develop protective immunity to PAM over successive pregnancies, and only primigravidae and secundigravidae present higher placental infection prevalence rates [24]. In *P. falciparum* transmission areas such as Benin, exposure is high and results in a fast acquisition of

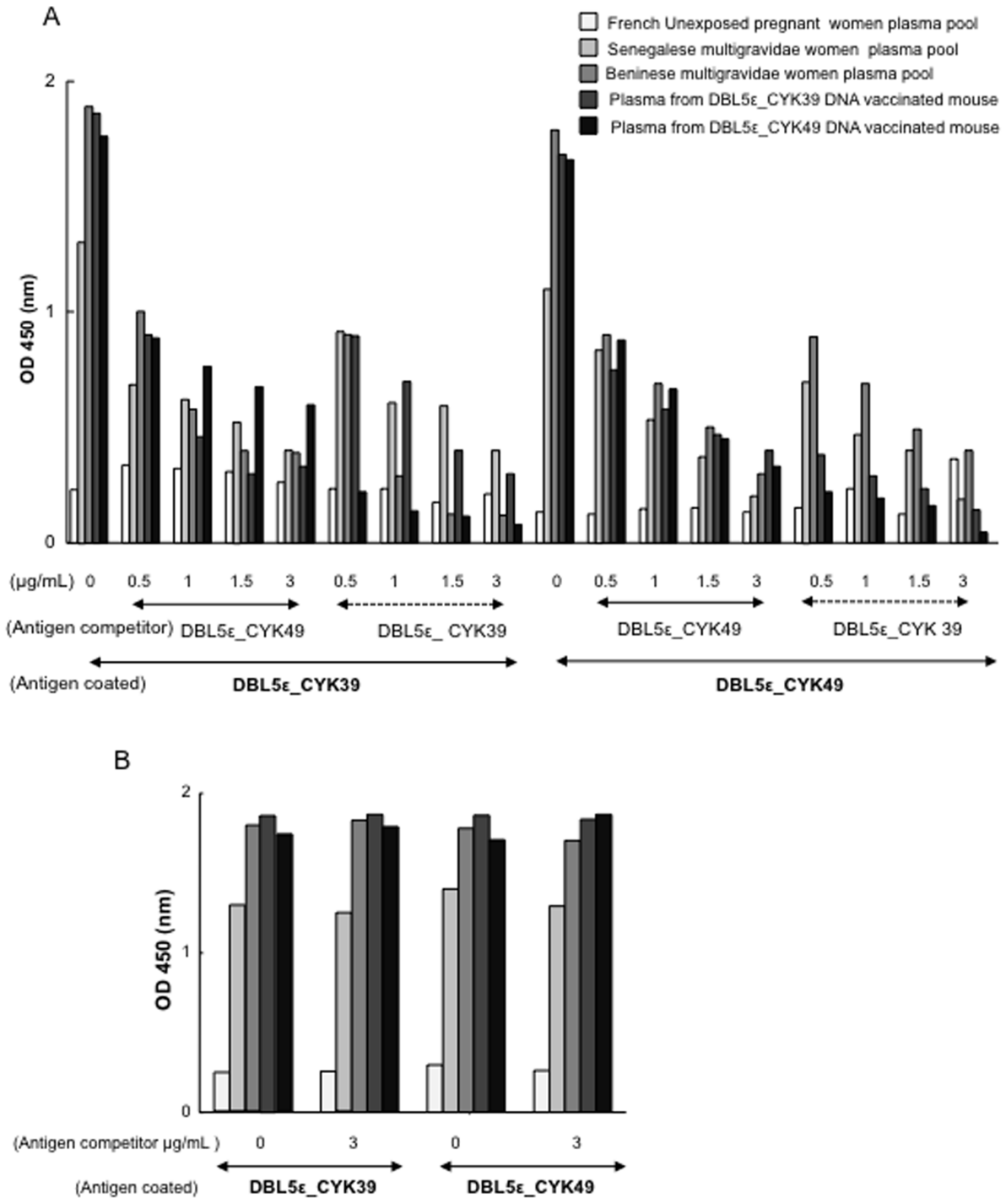


Figure 7. Cross-reactive antibody target between VAR2CSA DBL5ε variants. Cross-reactivity was determined by competition ELISA using either a multigravid plasma pool with high titer of VAR2CSA-specific antibodies (Beninese or Senegalese women), plasma from DBL5ε_CYK39 or CYK49 DNA genetic vaccinated mouse (A). NTS-DBL1α domain of VARO was used as negative control (B). Each colour shows the reactivity with the indicated antibodies.
doi:10.1371/journal.pone.0013105.g007

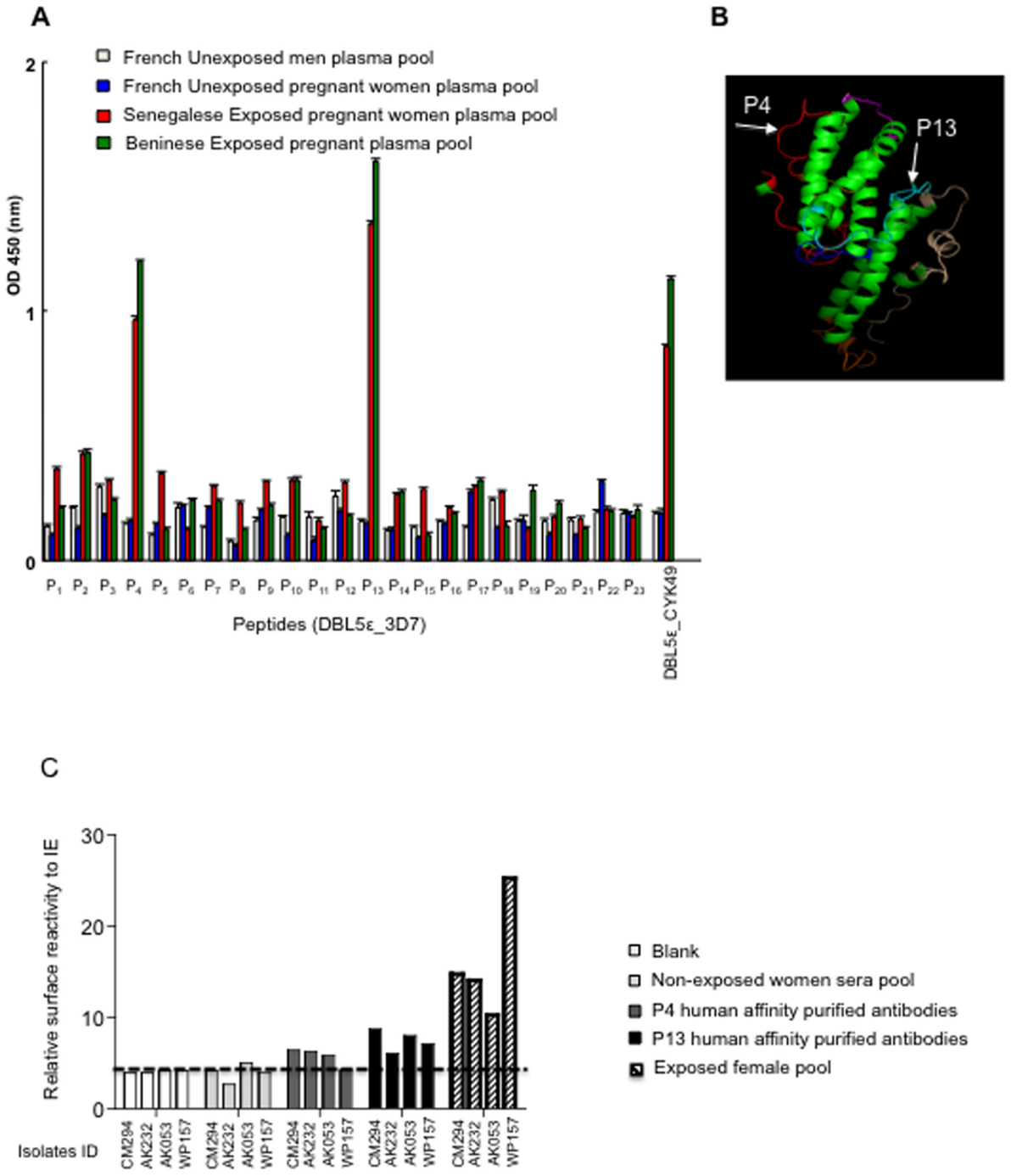


Figure 8. Reactivity of human specific conserved DBL5ε affinity purified antibodies with *P. falciparum* infected erythrocytes. (A): IgG recognition of 3D7-DBL5ε peptides library. (B): Mapping of P4 and P13 peptides on DBL5ε model [8]. (C): Senegalese women antibodies were affinity purified on peptides P4 and P13 and tested for reactivity against PAM Beninese parasite isolates. Flow cytometry analysis of human affinity-purified IgG against peptides P4 and P13 against PAM parasite isolates. Each colour shows the reactivity to native parasites with the indicated antibodies. Four isolates were tested with each IgG. Sample without primary antibody (blank), non-exposed women plasma pool, and exposed women plasma pool are used as control respectively.
doi:10.1371/journal.pone.0013105.g008

immunity while the acquisition may be delayed in areas of low and seasonal transmission such as in Senegal. In our two populations, multigravidae presented with higher antibody levels against VAR2CSA than primigravidae; but in Benin, where transmission is perennial, the mean antibody level was overall higher than that of women from Senegal. Among primigravidae, 57% of Beninese had anti-VAR2CSA antibodies at delivery compared to 16% of Senegalese. This could be explained by difference in malaria exposure in the study areas. A close comparison of the two VAR2CSA DBL5ε recombinant variants demonstrated that, despite a homology of 80% in their amino acid sequences, both variants presented some distinguishable characteristics. The DBL5ε_CYK39 exhibited a higher CSPG binding ability and a higher recognition by plasmas than the DBL5ε_CYK49 variant, although both constructs showed parity-dependent recognition patterns. This observation suggests that some variants can be more readily recognized than others. This can also be a useful consideration in vaccine development strategy, as not all VAR2CSA variants are likely to yield broad and high recognition or reactivity.

In the variable regions of DBL5ε distinct motifs were identified, the sero-reactivity of peptide containing TFKNI (P19) was assessed by ELISA. No reactivity was observed against this as shown in Figure 8A. Nevertheless, this result is not surprising as we clearly showed that TFKNI only were encountered in women presenting high parity status and may be expressed by uncommon variants. In the same effort to develop optimal VAR2CSA-based vaccine, it is advisable to target highly conserved residues or as many residues as possible that are accessible by host immune response to broaden the possibility of reaching all potential parasite populations expressing the VAR2CSA ligand. From the current observation it is obvious that like DBL3X, the DBL5ε domain variants share common and cross-reactive motifs. We identified two peptides (P4 and P13) in the highly conserved region of the DBL5ε domain that significantly reacted with plasma pool from pregnant women of different endemic areas. Affinity-purified antibodies against those peptides specifically reacted with placental parasites, confirming that these peptides are actually surface-exposed, as suggested by the 3D model. One such epitope in DBL5ε (peptide P63) was previously described which reacted strongly with Tanzanian female plasma [8]. DBL5ε peptide P4 identified in this study has 16 amino acids out of 20 in common with P63 peptide. Existence of such conserved and accessible epitopes supports the broad recognition observed on this particular DBL domain and emphasizes on its potential interest.

Knock-out studies have previously demonstrated the exclusive need for VAR2CSA to mediate IE binding to CSA [11], and it has been shown that four of the six Duffy-binding-like (DBL) domains of VAR2CSA individually have the ability to bind CSA *in vitro* [12,13,14,15,16], In this study, we confirmed the CSA-binding ability of recombinant DBL5ε to CSPG. Our results have demonstrated in our experimental conditions, that both placental isolate DBL5ε variants have certain affinity for CSPG. This result is in agreement with the fact that DBL5ε _CYK39 variant is able to bind to CSA and heparin sulfate [16]. However NTS-DBL1α domain of the VARO PfEMP1 that is not involved in the placental sequestration of parasites also presented a weak affinity to CSPG. The binding of VAR2CSA to placental CSPG plays a major role in malaria during pregnancy, and the understanding of this interaction will be valuable to define easily producible constructs that can induce adhesion inhibitory antibodies. Unlike CSA binding that is unique to PAM parasites, *in vitro* interaction of individual DBL with CSA is often seen with non-VAR2CSA DBLs. Whether such interactions of individual domain can predict for IE binding phenotype is debatable. Thus the CSPG interaction was used in the current study only as an analytic tool to characterize the properties of the both recombinant DBL5ε variants expressed. Recent studies have demonstrated that even though DBL3X and DBL6ε can bind to the same ligand, the sites of interaction differ in these domains [25,26]. Nevertheless, in each of these domains, the binding site involves residues that are conserved in parasite isolates from different geographic locations. We report in this study a difference in CSPG binding ability among two VAR2CSA DBL5ε variants. The structure of this domain has not yet been solved and residues which are essential for interaction are not identified.

In summary, we demonstrated for the first time that although VAR2CSA DBL5ε sequence has a limited antigenic diversity, it contains some molecular signatures that distinguish parasites according to the host parity. These findings have important implications for vaccine design based on VAR2CSA. Malaria-exposed women also develop antibodies against conserved parts of VAR2CSA DBL5ε domain. Two of such conserved epitopes were identified here and, naturally acquired antibodies to them stained native proteins on placental parasites. Our data support the importance of DBL5ε in the current effort of elucidating the parts of the VAR2CSA protein that can induce an antibody response with broad reactivity on placental parasites.

Materials and Methods

Parasite isolates

All *P. falciparum* PAM parasites for which sequences were generated were collected at delivery in a cross-sectional study conducted in Senegal in 2003[2]. Samples from 39 *P. falciparum* isolates were available for the study. The mean ± SD age of women who donated the parasites was 24±6.5 years. They were composed of 15 primigravidae, 6 secundigravidae, and 18 multigravidae. *P. falciparum* infected erythrocytes (IEs) were collected from parasitized placentas (parasite density ranging from 0.1% to 50%; mean ± SD,12.8±12.7) by flushing as previously described [2]. Collected IEs were conserved in Trizol LS (Invitrogen) and stored at −80°C until use. The binding ability of parasite isolates to CSA were evaluated [2]. Neonate birth weight was estimated by use of an electronic balance. There were 56% low birth weight LBW (<2500g) recorded.

Placental parasite "748" was collected in Tanzania, as described elsewhere [20].

Parasites used to evaluate antibody reactivity with the surface of IEs were freshly collected from pregnant women enrolled in the on going STOPPAM project based in the district of Comè, southwestern Benin [27].

Plasma samples collection

Plasma samples from malaria exposed women are from two different malaria endemic areas: Perennial (Benin) and seasonal (Senegal) *P. falciparum* transmission. Senegalese pregnant women were enrolled in a cohort study in 2001 in Thiadiaye [7]. Women presenting with fever and a positive blood smear were given curative treatment with chloroquine, the drug advocated in Senegal at the time of study for both prophylaxis and treatment.

In Benin, as described [28], pregnant women were enrolled in a cohort study conducted from July 2005 through April 2008 in Ouidah, a semirural town in Benin that is located 40 km west of Cotonou, the political capital of Benin. Perennial malaria transmission with seasonal peaks is mostly attributable to *P. falciparum* [29]. Sulfadoxine-pyrimethamine or mefloquine was given to women during the study.

Plasma samples from 24 French pregnant women and 8 adults men without *P. falciparum* exposure were used as negative controls.

All women plasma samples tested in this study were collected at delivery time.

Cloning and sequencing of placental var2csa DBL5ε genes

All VAR2CSA DBL5ε sequences were obtained from placental parasites complementary DNA (cDNA). Total RNA was extracted from parasites conserved in Trizol according to the manufacturer's instruction. The total RNA concentration was determined at 260 nm and RNA integrity was checked in 1% agarose gel. RNA samples were pretreated with DNAse I (Sigma-Aldrich). 5 U of RNase-free DNase per 5 ug of RNA was incubated at 37°C for 30 min, followed by 10 min heat inactivation at 65°C. All RNA samples were subsequently tested in real-time PCR for contamination with genomic DNA using a primer set for the housekeeping gene *seryl-tRNA synthetase*. cDNA was synthetised by reverse trancriptase (Superscript II, Invitrogen) and random hexamer primers, as described by the manufacturer. All VAR2CSA DBL5ε sequences were amplified using high fidelity enzyme (Phusion) with the following universal primers designed in highly conserved areas flanking the DBL5 DBL5ε+the hypervariable interdomain (Id5): DBL5ε Forward: 5′-GTC ACC CCC GGG GAC AAT GCA ATA AAA GAT TAC and DBL5ε Reverse: 5′-TAG GCA TTT GCG GCC GCC TTC AAG TTC AGC TGG AAT ATT. Two µl of cDNA was used for the PCR reactions. PCR products were inserted into a pAcGP67C Baculovirus Transfer Vector (BD). Ten to 15 colonies of each cloning were sequenced by GATC (www.gatc.com).

Cloning, expression and purification of recombinant VAR2CSA DBL5ε variants proteins

DBL5ε sequence from placental parasite isolate CYK 49 [2] was amplified from the corresponding cDNA with the following primers: 5′ ACT GGC AGG AAT TCA TGT TTG ATG ATC AGA CA and 3′ ATC GAC TGG CAG GCG GCC GCT TAA TGG TGA TGG TGA TGG TGT TTC ATA TCA TTA. PCR product was digested with *Eco*R1 and *Not*1 for cloning into the modified bacterial expression vector pET-21 (Novagen, http://www.novagen.com) to produce His-tagged recombinant proteins in *Rosetta gami* strain. The ligated vectors were transformed into *E. coli* DH5α strain, and positive clones were selected with ampicillin resistance. *Rosetta gami* cells transformed with recombinant plasmids, were cultured into LB broth containing ampicillin (50 µg/ml) at 30°C, and treated at the mid-log phase ($OD_{600} = 0.4$) with IPTG, to induce protein production. Cells were cultured at 25°C overnight, and harvested by centrifugation at 6,000 g at 4°C for 15 min. The pellet was washed, resuspended in cold buffer containing 10 mM Tris, 500 mM NaCl and protease inhibitor cocktail (Cocktail set N°III, Calbiochem), and sonicated. DBL5ε recombinant protein was purified from bacterial soluble fraction on Ni^{2+} metal-chelate agarose columns (GE Healthcare), and eluted with 10 mM Tris, 500 mM NaCl and 150 mM imidazole. Affinity chromatography step was followed by gel filtration. Recombinant DBL5ε protein from isolate CYK *39* [16] and NTS-DBL1α VARO [30,31] were produced, and purified under the same conditions.

Antibodies production

Specific antibodies to DBL5ε CYK39 or DBL5ε CYK49 were induced in mice by genetic immunization. Briefly, DNA injections were subcutaneously electro-transferred to 6-week-old Swiss female mice (Janvier, France) using 40 µg of plasmid DNA encoding either DBL5ε_CYK39 or DBL5ε_CYK49. All plasmids used for genetic vaccination are based on a pVax1 vector backbone (Invitrogen) in which the original cytomegalovirus (CMV) promoter has been replaced with the CMV promoter of the pCMVb plasmid (Clontech), as described [32]. Mice were electrotransferred on days 0, 21 and 45. Mice were bled before each electroporation, and a full bleed was collected 80 days (D80) after the first electroporation. Immune response was checked by ELISA on consecutive bleeds. All procedures complied with European and National regulations.

IgG from plasma of multigravidae living in an endemic area were purified on a Hi-Trap protein G HP column according to the manufacturer's recommendations (GE-Healthcare). The specificity of the purified antibodies was tested in ELISA against recombinant DBL5ε recombinant proteins (CYK39 and CYK49).

VAR2CSA proteins characterization by Western blotting

The soluble recombinant VAR2CSA DBL5ε proteins were checked by Sodium Dodecyl Sulfate-polyacrylamide gel electrophoresis and Western blotting. Protein samples (2–50 µg) were suspended in Laemmli-buffer (Tris/HCl 62.5mM, pH6.8, 2% SDS, 5% β-mercaptoethanol and 10% glycerol), subjected to SDS-PAGE [33] using a 4–12% acrylamide slab minigel (Invitrogen, Carisbad, CA, USA). Western blotting was performed with (2–30 µg) bacterial (induced, uninduced and nontransformed) lysates or purified eluates electrophoresed through 4–12% SDS-PAGE gels and electro-transferred to 0.2 µm Protan BA 83 nitrocellulose sheets (Schleicher & Schuell) for immunodetection. The membranes were blocked for 1 h with 5% nonfat dry milk in phosphate-buffered saline (PBS) with 0.1% Tween® and then incubated separately with either a 1:5000 dilution of a monoclonal anti-histidine HRP conjugated antibody (46-0707, Invitrogen) or a 1:1000 dilution of DBL5ε_CYK39 or DBL5ε_CYK49 antiserum from vaccinated mouse D50 (day 50) or 1:1000 of IgG purified from plasma of multigravidae living in an endemic area. Immune complexes were detected with a HRP coupled with either anti-mouse IgG antibody (1:10 000, AP127P Sigma-Aldrich) or anti-human IgG antibody (1:10 000, A0170 Sigma-Aldrich).

Competition ELISA, peptide ELISA and affinity purification of antibodies

Prior to competition ELISA, both VAR2CSA DBL5ε constructs were used to assess the plasma levels of anti VAR2CSA IgG of 160 malaria exposed pregnant women from Benin (primigravidae n = 80, multigravidae n = 80) and Senegal (primigravidae n = 24; multigravidae n = 50), French unexposed pregnant women (n = 16), and French unexposed men (n = 8). ELISA was carried out on plates coated with 0.5 µg/ml of the DBL5ε. The IgG plasma levels were expressed as Optical densities (OD) values read at 450nm. A pool of plasma samples from unexposed French pregnant women was used as a negative control whereas a pool of plasma samples from multigravidae pregnant Senegalese women, previously demonstrated to have high levels of anti-VSA IgG (VSA: Variant surface antigen) against placental isolates, was used as a positive control.

For competition ELISA, microtiter plates (Nunc 442404) were coated with each antigen (DBL5ε_CYK39, DBL5ε_CYK49, NTS-DBL1α-VARO, 0.5 µg/ml in PBS). Five different plasma pools were individually pre-incubated for 2 h at room temperature (RT) with increasing concentrations of competing antigen (0.5, 1, 1.5, and 3 µg/ml): Beninese pregnant women plasma pool (diluted 1:500), Senegalese pregnant women plasma pool (diluted 1:500), DBL5ε_CYK39 plasma from DNA vaccinated D50 (1:100 000),

DBL5ε_CYK49 plasma from DNA vaccinated D50 (1:40 000), and French unexposed women plasma pool (1:100). After incubating the plates with blocking buffer (PBS, 0.5 M NaCl, 1% Triton X-100, 1% BSA) for 2 h at RT, the pre-absorbed pool were added to the antigen-coated wells in duplicate and incubated overnight at 4°C. In addition to the pre-absorbed plasma pool, a non-absorbed pool was included for each coating antigen. Following washing of the plates four times with washing buffer (PBS, 0.5 M NaCl, 1% Triton X-100, pH 7.4), the secondary antibody (Goat anti-human IgG HRP, A0170, Sigma-Aldrich for human plasma and Goat anti-mouse IgG HRP, AP127P, Chemicon) diluted 1:4000 in blocking buffer was added, and incubated for 1 h at RT. Plates were washed four times, and antibody reactivity visualized by the addition of TMB (Tetramethylbenzidine). Coloured reactions were stopped by the addition of 0.5 M H_2SO_4 and OD was measured at 450 nm.

Peptides and antibodies affinity purification of antibodies

DBL5ε of 3D7 PFL0030c *var2csa* sequence (Genbank accession number. XM_001350379) was used to design peptides. A library of 23 peptides (70% purity) each consisting of 20 amino acids and having an overlap of 6 amino acids was synthesized (Sigma Genosys). All peptides had a free amine at the N- and a free acid at the OH-terminus. ELISA was carried out on plates coated with 5 µg/ml of each peptide. VAR2CSA antibodies reactivity against those peptides was measured using Senegalese pregnant women plasma pool 1:100 (pool was obtained with n = 30 multigravidae plasma) and Beninese pregnant women plasma pool 1:100 (pool was obtained with n = 30 multigravidae plasma). Plasma samples from Unexposed French men (n = 8) and pregnant women (n = 16) were used as negative controls.

The two peptides (P4: $_{2037}$RRQLCFSRIVRGPANLRNLK$_{2056}$ and P13: $_{2163}$SWCTIPTTETPPQFLRWIKE$_{2182}$) which reacted with malaria exposed pregnant women plasma pool were used for affinity purification of antibodies. This was done using HiTrap NHS-activated HP columns (GE Healthcare) according to the manufacturer's instructions. In brief, 1 mg of each synthetic peptide was dissolved in coupling buffer 0.2 M $NaHCO_3$, 0.5 M NaCl (pH 8.3), and applied to the 1 ml column previously equilibrated with 3×2 ml of ice-cold 1 mM HCl. After coupling, the columns were washed alternating 0.5 M ethanolamine, 0.5 M NaCl (pH 8.3) and 0.1 M acetate, 0.5 M NaCl (pH 4), followed by a final wash with PBS (pH 7.4). One ml of Senegalese pregnant women plasma pool was diluted in PBS (1:1), filtered through a 0.45-µm filter and applied to the column at a flow rate of 1 ml / min. After washing the column in 7 ml PBS, affinity-bound antibodies were eluted in fractions with a total volume of 3 ml of 0.1 M glycine-HCl (pH 2.8) and neutralized in 1 M Tris (pH 8). The specificity of the purified antibodies was tested in ELISA against the peptide used for affinity purification.

Antibody recognition of surface VAR2CSA

P. falciparum-IEs collected *ex vivo* from the placenta of Beninese women were used without additional *in vitro* culture. Flow cytometry was used to test the reactivity of the antibodies against either the P4 or P13 peptides with parasite isolates, as described elsewhere [34]. Briefly, mature parasites (four placental isolates) were enriched to contain >75% PE at late-stage trophozoite and schizont stages by exposure to a strong magnetic field. Aliquots of $\sim 2\times10^5$ PE were labeled by ethidium bromide and sequentially exposed to 20 µl human purified IgG (~0.2 µg IgG) and 1 µl goat anti-human IgG-FITC (Sigma). Data were acquired using FACS Calibur (BD Biosciences, Franklin 10 Lakes, NJ). All samples relating to a particular parasite isolate were processed and analyzed in a single assay.

Interaction properties of the recombinant DBL5ε proteins

Binding to CSPG (decorin D8428, Sigma-Aldrich) was performed mainly as described elsewhere [35]. Briefly falcon plates (351172, Becton Dickinson) were coated with either 5 µg/ml of CSPG in PBS or with 1% BSA in PBS for background measurement (overnight at 4°C). Following coating, the wells were blocked with TSM binding buffer (20 mM Tris–HCl, 90 mM NaCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 0.05% Tween-20 and 1% BSA, pH 7.4 at 25°C) at room temperature for 6h. A dilution series (0.4–40 µg/ml) of the DBL5ε recombinant domains in TSM binding buffer was added in each well and incubated overnight at 4°C with gentle shaking. After washing three times in TSM washing buffer, 100 µl of anti-His tag-HRP antibody diluted 1:3 000 in binding buffer was added to each well and incubated for 2h at room temperature. The assay was finalised with three washes and developed using 100 µl per well of TMB substrate for 30 min. Absorbance was measured at 450 nm after quenching the reaction with 100 µl of 0.5 M H_2SO_4.

Inhibition of recombinant domain binding to CSPG was performed mainly as the above described ELISA analysis, but using a constant protein concentration (5 µg/ml) pre-mixed with increasing amounts of soluble CSA (0.5–400 µg/ml).

In silico analyses of VAR2CSA sequences from field isolates

Multiple alignment. Initially a master data file was created, containing sequence ids, experimental parameters (where available) and unaligned sequences. The DBL5ε were aligned using ClustalW2 [36] with default options. The resulting alignment was inspected and manually adjusted. Aligned sequences were then inserted in the master file.

Evaluation of system diversity by calculation of Shannon entropy. The Shannon entropy [37] was calculated for each position in the multiple alignment as:

$$H(p) = -\sum_a p_a \log_2(p_a)$$

Briefly on values of H: $H = 0$: Complete conservation, only one residue present at the given position. $0 < H \leq 1$: Considered highly conserved. $1 < H \leq 2$: Considered conserved. $2 < H \leq 4.3$ considered variable. The calculated Shannon Entropy per multiple alignment position was subsequently depicted.

Homology modeling. DBL5ε homology models were created by submitting the multiple alignment to the HHpred server [38]. Best hit was chosen based on an evaluation of score and structure resolution (VAR2CSA DBL3x domain, PDB ID: 3bqk) [26]. One primi- and one multigravidae representative sequence were chosen and submitted individually to HHpred. The resulting models were loaded into PyMOL [39] and aligned for visual analysis of structural impact of motifs. The models were validated by submission to the ProQ server [40]. Likewise was a model of DBL5ε-3d7 created for mapping purposes.

Mapping of sequence variability. The sequence variability was mapped onto a homology model of DBL5ε-3d7 by submission to the H2PDB server [41]. The resulting pdb-file was loaded into PyMOL and variability was visualised by heat-map colouring (colour by b-factor).

SigniSite analysis. A statistical *In silico* analysis of the multiple alignment was performed using the SigniSite server [23]. Briefly: The SigniSite server performs a non-parametric

statistical evaluation of the distribution of each residue at each position, aiming at identifying any significant association with a sequence associated numerical parameter, specified at submission. As a prerequisite for submission to SigniSite is the association of a numerical parameter to each sequence, sequence files were created for each numerical parameter containing the DBL5ε sequences and the associated numerical parameter (where available). Numerical isolate parameters were: Maternal age at delivery [year(s)], Concentration of parasites in peripheral blood of the mother [/μL], Concentration of parasites in the placental blood [/μL], Parity, Birth weight [g], CSA binding density [mean/mm^2], CSPG binding density [mean/mm^2]. Some of the women were infected with more than one parasite and thus some isolates contain more than one sequence. It should be noted that (i) numerical values associated with a particular isolate were assigned to all the sequences identified in that particular isolate and (ii) not all parameters were available for all sequences, if no parameter was available, the sequence was excluded from evaluation. As SigniSite performs multiple testing, it was imperative to reduce the number of tests performed prior to submission. This was done in two steps: (i) Exclusion of all positions in the multiple alignment with $H = 0$ (If just one residue is present at a given position, no significant distribution is possible). (ii) Evaluating only the top 15% most variable positions as estimated by the entropy calculation (It is more likely to identify a significant distribution at the most variable positions). Following this, the before mentioned sequence files were reduced to only contain the positions selected for testing. The sequence files were subsequently submitted to evaluation by SigniSite with the following settings: Significance threshold = 0.05, Correction for multiple testing using the Bonferroni single-step, Consider values given in fasta header and Choose decreasing order. The normal distributed Z-scores were converted to p-values by standard method.

Statistical analysis

Comparison of anti-VAR2CSA antibodies levels between groups was tested by nonparametric Mann-Whitney test. Correlations were examined by use of Pearson's test. The chi^2 test was used to examine differences between categorical variables. The Fisher's exact test was used to evaluate significance when analysing motifs and parasite expressing specific motifs identified. The significance limit was $P<0.05$. When evaluating DBL5ε sequences containing Q_{303} vs. E/K_{303}, population means, with respect to placental parasites CSA/CSPG binding, were calculated and a two sample t-test was applied to test if differences in means were significant ($p<0.01$).

Supporting Information

Figure S1 Multiple alignment of parasite isolates VAR2CSA DBL5ε sequences. cDNA from 40 placental parasites isolates (39 placental isolates from Senegal and one from Tanzania) were amplified, cloned, and sequenced. Sequence ids are given at the far left. The Tanzanian isolate was isolate 748 (sequences 748_1/2a and 748_1/2b) corresponding to the DBL5ε domain amplified in this isolate. The remaining sequences correspond to those obtained in isolates from Senegal. The remaining CYK are Senegalese isolates. The CYK suffix corresponds to the placenta id from which the isolate was extracted. The DBL5ε and ID5 highly conserved (blue, Shannon entropy $0 \leq H \leq 1$), conserved (green, $1 < H < 1.5$), and relatively variable (red, $1.5 \leq H \leq 2$) blocks, are indicated. The 15% most variable positions were selected and marked with "x".
Found at: doi:10.1371/journal.pone.0013105.s001 (0.11 MB PDF)

Acknowledgments

We thank Gwladys Bertin, Nadine Fievet, Achille Massougbodji, Alioune Gaye for parasite and plasma collection, and Alexandre Juillerat for providing recombinant NTS-DBL1α domain of VARO.

Author Contributions

Conceived and designed the experiments: SG OL PD NTN. Performed the experiments: SG LJ CE CT NTN. Analyzed the data: SG LJ OL PD NTN. Contributed reagents/materials/analysis tools: SG MQ JG PD. Wrote the paper: SG LJ NTN.

References

1. Duffy PE, Fried M (2003) Antibodies that inhibit Plasmodium falciparum adhesion to chondroitin sulfate A are associated with increased birth weight and the gestational age of newborns. Infect Immun 71: 6620–6623.
2. Tuikue Ndam NG, Fievet N, Bertin G, Cottrell G, Gaye A, et al. (2004) Variable adhesion abilities and overlapping antigenic properties in placental Plasmodium falciparum isolates. J Infect Dis 190: 2001–2009.
3. Beeson JG, Rogerson SJ, Cooke BM, Reeder JC, Chai W, et al. (2000) Adhesion of Plasmodium falciparum-infected erythrocytes to hyaluronic acid in placental malaria. Nat Med 6: 86–90.
4. Fried M, Duffy PE (1996) Adherence of Plasmodium falciparum to chondroitin sulfate A in the human placenta. Science 272: 1502–1504.
5. Fried M, Nosten F, Brockman A, Brabin BJ, Duffy PE (1998) Maternal antibodies block malaria. Nature 395: 851–852.
6. Salanti A, Dahlback M, Turner L, Nielsen MA, Barfod L, et al. (2004) Evidence for the involvement of VAR2CSA in pregnancy-associated malaria. J Exp Med 200: 1197–1203.
7. Tuikue Ndam NG, Salanti A, Le-Hesran JY, Cottrell G, Fievet N, et al. (2006) Dynamics of anti-VAR2CSA immunoglobulin G response in a cohort of senegalese pregnant women. J Infect Dis 193: 713–720.
8. Andersen P, Nielsen MA, Resende M, Rask TS, Dahlback M, et al. (2008) Structural insight into epitopes in the pregnancy-associated malaria protein VAR2CSA. PLoS Pathog 4: e42.
9. Salanti A, Staalsoe T, Lavstsen T, Jensen AT, Sowa MP, et al. (2003) Selective upregulation of a single distinctly structured var gene in chondroitin sulphate A-adhering Plasmodium falciparum involved in pregnancy-associated malaria. Mol Microbiol 49: 179–191.
10. Bengtsson D, Sowa KM, Salanti A, Jensen AT, Joergensen L, et al. (2008) A method for visualizing surface-exposed and internal PfEMP1 adhesion antigens in Plasmodium falciparum infected erythrocytes. Malar J 7: 101.
11. Viebig NK, Levin E, Dechavanne S, Rogerson SJ, Gysin J, et al. (2007) Disruption of var2csa gene impairs placental malaria associated adhesion phenotype. PLoS One 2: e910.
12. Avril M GB, Lépolard C, Viaud N, Scherf A, Gysin J (2006) Characterization of anti-var2CSA-PfEMP1 cytoadhesion inhibitory mouse monoclonal antibodies. Microbes Infect 8: 2863–2871.
13. Gamain B, Trimnell AR, Scheidig C, Scherf A, Miller LH, et al. (2005) Identification of multiple chondroitin sulfate A (CSA)-binding domains in the var2CSA gene transcribed in CSA-binding parasites. J Infect Dis 191: 1010–1013.
14. Resende M, Nielsen MA, Dahlback M, Ditlev SB, Andersen P, et al. (2008) Identification of glycosaminoglycan binding regions in the Plasmodium falciparum encoded placental sequestration ligand, VAR2CSA. Malar J 7: 104.
15. Dahlback M, Rask TS, Andersen PH, Nielsen MA, Ndam NT, et al. (2006) Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in P. falciparum placental sequestration. PLoS Pathog 2: e124.
16. Gangnard S, Ndam N, Gnidehou S, Quiviger M, Juillerat A, et al. (2010) Functional and immunological characterization of the var2CSA-DBL5varepsilon domain of a placental Plasmodium falciparum isolate. Mol Biochem Parasitol.
17. Trimnell AR, Kraemer SM, Mukherjee S, Phippard DJ, Janes JH, et al. (2006) Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. Mol Biochem Parasitol 148: 169–180.
18. Guitard JAP, Ermont C, Gnidehou S, Fievet N, Lund O, et al. (2010) Plasmodium falciparum population dynamics in a cohort of pregnant women in Senegal. Malar J 9(1): 165.
19. Nielsen MA (2007) Plasmodium falciparum: VAR2CSA expressed during pregnancy-associated malaria is partially resistant to proteolytic cleavage by trypsin. Experimental Parasitology 117: 1–8.

20. Magistrado P, Salanti A, Tuikue Ndam NG, Mwakalinga SB, Resende M, et al. (2008) VAR2CSA expression on the surface of placenta-derived Plasmodium falciparum-infected erythrocytes. J Infect Dis 198: 1071–1074.
21. Barfod L (2007) Human pregnancy-associated malaria-specific B cells target polymorphic, conformational epitopes in VAR2CSA. Molecular Microbiology 63(2): 335–347.
22. Sander AF, Salanti A, Lavstsen T, Nielsen MA, Magistrado P, et al. (2009) Multiple var2csa-type PfEMP1 genes located at different chromosomal loci occur in many Plasmodium falciparum isolates. PLoS One 4: e6667.
23. Hoof I (2009) Prediction and analysis of MHC class I binding specificities beyond humans. PhD thesis, center for biological Sequence Analysis Department of Systems Biology technical University of Denmark.
24. Steketee RW, Breman JG, Paluku KM, Moore M, Roy J, et al. (1988) Malaria infection in pregnant women in Zaire: the effects and the potential for intervention. Ann Trop Med Parasitol 82: 113–120.
25. Khunrae P, Philip JM, Bull DR, Higgins MK (2009) Structural comparison of two CSPG-binding DBL domains from the VAR2CSA protein important in malaria during pregnancy. J Mol Biol 393: 202–213.
26. Higgins MK (2008) The structure of a chondroitin sulfate-binding domain important in placental malaria. J Biol Chem 283: 21842–21846.
27. Yadouleton AWPG, Asidi A, Moiroux N, Bio-Banganna S, Corbel V, et al. (2010) Insecticide resistance status in Anopheles gambiae in southern Benin. Malar J 9: 83.
28. Briand V, Bottero J, Noel H, Masse V, Cordel H, et al. (2009) Intermittent treatment for the prevention of malaria during pregnancy in Benin: a randomized, open-label equivalence trial comparing sulfadoxine-pyrimethamine with mefloquine. J Infect Dis 200: 991–1001.
29. Akogbeto M, Modiano D, Bosman A (1992) Malaria transmission in the lagoon area of Cotonou, Benin. Parassitologia 34: 147–154.
30. Juillerat A, Igonet S, Vigan-Womas I, Guillotte M, Gangnard S, et al. (2010) Biochemical and biophysical characterisation of DBL1alpha1-varO, the rosetting domain of PfEMP1 from the VarO line of Plasmodium falciparum. Mol Biochem Parasitol 170: 84–92.
31. Vigan-Womas I, Guillotte M, Le Scanf C, Igonet S, Petres S, et al. (2008) An in vivo and in vitro model of Plasmodium falciparum rosetting and autoagglutination mediated by varO, a group A var gene encoding a frequent serotype. Infect Immun 76: 5565–5580.
32. Leblond J, Mignet N, Largeau C, Spanedda MV, Seguin J, et al. (2007) Lipopolythioureas: a new non-cationic system for gene transfer. Bioconjug Chem 18: 484–493.
33. Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227: 680–685.
34. Flick K (2001) Role of nonimmune IgG bound to PfEMP1 in placental malaria. Science 293: 2098–2100.
35. Resende M, Ditlev SB, Nielsen MA, Bodevin S, Bruun S, et al. (2009) Chondroitin sulphate A (CSA)-binding of single recombinant Duffy-binding-like domains is not restricted to Plasmodium falciparum Erythrocyte Membrane Protein 1 expressed by CSA-binding parasites. Int J Parasitol 39: 1195–1204.
36. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
37. Shannon CW, W (1949) The mathematical Theory of Communication; Press UoI, editor.
38. Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33: W244–248.
39. Delano W (2002) The Pymol Molecular Graphics System.
40. Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12: 1073–1086.
41. H2PDB server. Available: http://bio.dfci.harvard.edu/Tools/entropy2pdb.htm.

Part IV

Development and Application of Bioinformatics tool for Signal Detection in High Throughput, High Density Peptide Microarray

The ability to speak does not make you intelligent.

Qui-Gon Jinn

4

VAR2CSA linear B-cell epitope discovery

4.1 Introduction

Immunity to placental malaria is gradually acquired and mediated by the accumulation of antibodies capable of blocking the VAR2CSA:receptor interaction. Obtaining a deep understanding of the immunogenicity of VAR2CSA can facilitate not only the identification of epitopes capable of inducing blocking antibodies, but also the immunodominant epitopes acting as an immunological smoke-screen. The purpose of these epitopes are to divert the 'attention' of the human immune system, away from blocking antibody inducing epitopes and thusly maintaining the cytoadherence capability. These epitopes are naturally to be avoided in the context of vaccine development. The identification of which epitopes are capable of inducing high titer blocking antibodies response is paramount in the search for a vaccine against placental malaria. The purpose of

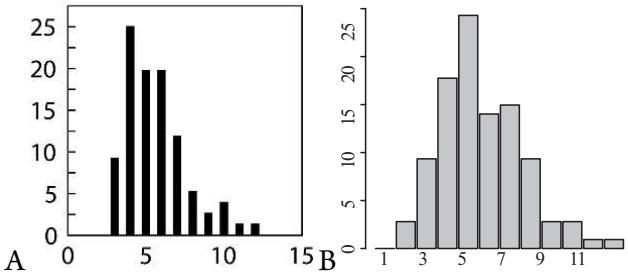


Figure 4.1.1: A: 2006 structural analysis of antibody:antigen complexes in the Protein Data Bank (PDB) [66] **B**: Likewise, but from 2013 [87]. Axis annotation: x-axis it the length of consecutive amino acid residues; y-axis it the proportion of epitopes found to have that length.

the the High Density Peptide Microarray (HDPMa) is to provide a high-throughput technology, with the capability of screen large libraries of peptides for an activity of choice. As with all high-throughput methods, the downstream challenge is how to handle the large amounts of data being generated. This challenge is the offset for the work described in this part of the thesis. The aim of this study was a 2-step process:

1. *Develop a statistical robust method for signal detection in High Throughput, High Density Peptide Microarray*

2. *Once developed, apply to VAR2CSA linear B-cell epitope discovery*

An initial concern was that the consensus regarding VAR2CSA epitopes is that the majority are conformational [41] and since it has been proposed that more than 90% of epitopes are conformational [10, 150], this technology may at first not seem suitable for discovery of linear b-cell epitopes. However Structural studies of epitope:antigen interactions, have shown that epitopes consists of a set of contact-points, which combined with a linear stretch of amino acid residues, acting as an antigenic determinant, constitute the epitope [66, 87]. The length of the linear determinant ranges from 4 to 7 consecutive amino acid residues (See fig. 4.1.1).

4.2 Materials

4.2.1 Generation of 15-mers from reference sequences

VAR2CSA derived peptides were created using a 'sliding-window' of 15 residues. The reference sequences were from standard strains 3D7 ($n_{res} = 2,671$) and FCR3 ($n_{res} = 2,715$), both with n-terminal and c-terminal spacer 'GAGAGAGAGAGAGAG'. The purpose of the spacer was to increase the availability of the terminus regions. Furthermore the FCR3 sequence had the V5 epitope 'GKPIPNPLLGLDST' incorporated in the c-terminal, along with a his-tag 'HHHHHH' (FCR3 and 3D7 sequence are available in supplementary materials). The V5 tag was used for positive control. Table 4.2.1 summarises 15-mer counts in the two reference sequences.

Strain	$n_{15-mers}$	$n_{15-mers}^{unique}$
FCR3	2,701	2,700
3D7	2,657	2,656
Pool	5,358	4,731

Table 4.2.1: Overview of sequence data.

4.2.2 HDPMa incubation

The chip used in this experiment, was subdivided into 24 sectors, each separated by highly hydrophobic teflon-barriers, allowing us to conduct 24 separate incubations based on the immunisations summarised in table 4.2.2.

4.2.3 HDPMa output

The output generated from reading the chip is a tab-separated text file with a total of 134,647 lines. From this file 105,141 15-mer peptides can be extracted each of which has an associated signal-to-noise (S/N) ratio and one of 24 sectors. Within the total set of 15-mers 3,748 are unique.

S	Immunisation
1	Full length FCR3
2	Full length 3d7 (Without v5 epitope)
3	Full length 3d7 (With v5 epitope)
4	Full length FCR3 (As 1 but from another animal and IgG purified)
5	Full length FCR3 purified on heterologous 3D7 Full length column
6	Full length 3D7 purified on homologous 3D7 Full length column
7	Full length 3D7 purified on heterologous FCR3 Full length column
8	Full length FCR3 purified on all single domains
9	DBL4 FCR3 vaccine construct (Not blocking)
10	DBL4 3D7 vaccine construct (Not particular blocking)
11	DBL4 FCR3 vaccine construct (blocking)
12	DBL4 from placental isolate 4708 (Cannot be inhibited by FCR3 generated sera)
13	DBL4 FCR3 vaccine construct produced in coli
14	DBL4 FCR3 vaccine construct produced in coli c-terminal truncated DBL4
15	DBL4 FCR3 vaccine construct produced in coli c-terminal truncated DBL4
16	Full length FCR3 immunised and boosted with FCR3 DBL4
17	FCR3 DBL4 vaccine and boosted with Full length FCR3
18	FCR3 DBL4 + Full length FCR3 vaccinated
19	VAR 1 (irrelevant protein) control
20	DBL1-2 vaccine FCR3
21	DBL4 vaccine FCR3 c-terminal truncated
22	DBL1-3 vaccine FCR3
23	DBL1-4 vaccine FCR3
24	Negative control (May lack inhibition from serum protein)

Table 4.2.2: Overview of immunisations. S refers to one of the 24 sectors on the HDPMa chip. DBL refers to one of the six VAR2CSA domains, FCR3 and 3D7 are standard VAR2CSA sequences, the V5 epitope, `GKPIPNPLLGLDST`, is used as positive control.

The discrepancy between the 4,731 possible 15-mers from FCR3 and 3D7 and the 3,748 unique 15-mers in the data from the read chip is due to some of the fields having been marked as 'excluded'. These fields are excluded because they could not be read properly, due to e.g. a spec of dust or similar. This of course means that not every unique FCR3/3D7 15-mer is covered in the data set, however since we are using a sliding-window approach, each position in FCR3 and 3D7 are covered by at least one 15-mer.

4.3 Methods

4.3.1 Direct signal mapping with standard score normalisation

Initially we applied the naive method op simply mapping the average signal to each VAR2CSA position as follows:
For each of the 24 sectors

1. Extract all 15-mers and associated signals from the raw data file
2. Assign mean signal to any non-unique 15-mers
3. Map 15-mers and associated signal onto reference sequence
4. Assign the mean signal s to each p_i in the reference sequence
5. Standard score normalise (SSN), such that $z(s_{p_i}) = \frac{s_{p_i} - \mu_{sec}}{\sigma_{sec}}$
6. Assume that the signal distribution in each sector can be approximated by the normal distribution and assign *p-values* under $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$
7. Use standard Bonferroni-single-step correction for multiple testing, counting each position as a test

Where the null-hypothesis is that there is an equal amount of signal and noise, i.e., the ratio between signal and noise is 1. SSN is performed on a set of values, by taking each individual value, subtracting the mean of the set and dividing by the standard score. After SSN, the mean of the set will be 0 and the standard deviation will be 0. Therefore the after SSN equal values of signal and noise will result in a SSN S/N value of '0'. Following this epitopes are identified as continuos stretches of residues, for which $z(s_{p_i}) > z_{1-\alpha/(2n_{tests})}$

4.3.2 Identification of epitope linear determinant using a *k-mer* based non-parametric approach

Based on the previous mentioned structural studies of antibody:antigen interaction and subsequent identification of linear determinant, the aim of this approach was to identify linear determinants, by applying a *k-mer* sub-division of the 15-mers and subsequently performing a statistical evaluation of the resulting *k-mer* population. Using the *SigniSite* method as inspiration, we turned to the non-parametric Wilcoxon Rank-Sum test [4, 127]. The method described here can be viewed as an expansion of the *SigniSite* method from '1-mers' to 'k-mers'. The main difference being that *SigniSite* makes positional evaluations based on the mean ranks, whereas this method evaluated all possible k-mers using the sum of the ranks. The method was implemented as follows:
For each of the 24 sectors

1. Extract all 15-mers and associated signal from the raw data file
2. Subdivide each 15-mer into $(15 - k + 1)$ k-mers, each associated with the same signal as the 15-mer from which they originated
3. Sort the generated k-mers, with respect to descending signal
4. Rank the sorted k-mers, such that the *k-mer* with the highest signal gets a rank of 1, the second highest signal 2 and so on. In case of tied values, each tied *k-mer* gets a rank corresponding to the average of the ranks, the tied k-mers would occupy, were they not tied.
5. For each unique k-mer, compute the sum of the ranks, $R1$

6. For each $R1$, compute $E(R1)$ and $Var(R1)$ as follows:

$$E(R1) = \frac{1}{2} \cdot n_1(N+1) \tag{4.1}$$

$$Var(R1) = \frac{n_1 n_2 (N+1)}{12} \tag{4.2}$$

$$Var(R1) = \frac{n_1 n_2}{12N(N-1)} \left[N^3 - N - \sum_{i=1}^{n} (t_i^3 - t_i) \right] \tag{4.3}$$

   Where $E(R1)$ is the expected mean, $Var(R1)$ is the variance in the absence and presence respectively of tied values. n_1 is the number of the unique k-mer, which is currently being tested, n_2 is the number of all other k-mers and $N = n_1 + n_2$. t_i is the number of tied values in the *i'th* of n tied group, e.g. given ranks $(1, 1, 3, 3, 5, 6, 7, 7, 7, 10)$, the tiesum, $\sum_{i=1}^{n}(t_i^3 - t_i)$, is $(2^3 - 2) + (2^3 - 2) + (3^3 - 3) = 36$

7. Lastly compute the test statistic T

$$T = q \cdot \frac{|R1 - E(R1)| - \frac{1}{2}}{\sqrt{Var(R1)}} \tag{4.4}$$

   Where $q = -1$ if $R1 > E(R1)$ and $q = 1$ otherwise. A negative T thus implies that $R1$ is higher than expected and that the association therefore is with high ranks, which is equivalent with 'weak' values and vice versa.

Henceforth the value of the computed T statistics will referred to as the *z-score* of the k-mer. Once a list of *k-mers* and associated *z-scores* have been computed, the *k-mers* can be mapped onto reference sequences FCR3 and 3D7. Each position in the reference sequence is assigned the *z-scores* of the *k-mer* starting at this p_i. The threshold for significance is adjusted using standard 'Bonferroni-single-step' correction for multiple testing, such that $z_{adj} = z_{1-\alpha/n_{tests}}$, where n_{tests} is the number of unique mapped *k-mers* and α is the level of significance usually 5%. Here we apply a one-sided test, since we are interested in identifying k-mers which are significantly associated with a high signal, i.e. $z > 0$. Epitopes are

identified by positional traversing the mapped *z-scores* and identifying continuous stretches of k-mers, for which $z > z_{adj}$. Each identified stretch of residues can then be ranked according to the *k-mer* with the highest *z-score* within the stretch. Furthermore this *k-mer* is recorded as the linear determinant, in that it contributes the largest signal with the stretch. Based on the findings summarised in fig. 4.1.1, we chose $k = 5$ for this analysis. Table 4.3.1 summarises the 5-mer population, which can be generated from reference sequences FCR3 and 3D7. Prior to applying the above described *k-mer* method, we filtered the

Strain	n_{5-mers}	n_{5-mers}^{unique}
FCR3	2,711	2,667
3D7	2,667	2,627
Pool	5,378	3,923

Table 4.3.1: Overview of sequence data.

raw data. Of a total of 105,141 15-mers, 64,413 had an associated signal-to-noise ratio of 1 or smaller. In other words $61,26\%$ of the data contained as much or more noise than actual signal. We therefore chose only to include 15-mers with an associated signal-to-noise ratio of more than 1.0. Considering that 60,168 15-mers (3,659 unique), equivalent to $57,23\%$ of the data, had an associated signal-to-noise ratio of exactly 1.0, this also made intuitively sense in that no information can be obtained from such a large set of quantitatively inseparable peptides. Filtering away these data-point, which contain no information, we end up with a total of 40,728 15-mers of which 3572 are unique.

4.4 Results

4.4.1 Descriptive statistics of empirical signal-to-noise ratios

Fig. 4.4.1A depicts the empirical cumulative distribution function (CDF) of a total of $105,141$ HDPMa S/N ratios, with a mean of 1.062 and a standard deviation of 0.165. The minimum value is 0.2 and the maximum is 6.9. The CDF

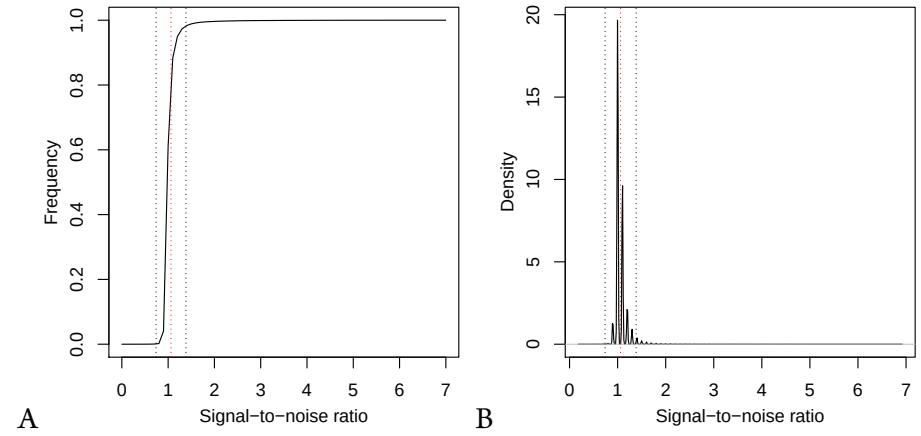


Figure 4.4.1: A: Empirical Cumulative Distribution function and **B**: emperical density for a total of 105,141 HDPMa signal-to-noise ratios. The dotted red vertical lines are $\mu = 1.062$ and the two vertical black dotted lines are $\mu \pm 1.96 \cdot \sigma$, where $\sigma = 0.165$. The plot clearly illustrate how the bulk of the data is centred around a signal-to-noise-ratio of 1.

clearly illustrates how the majority of the signals are centred around $S/N = 1$. The precision of the read signals are only to the first decimal, the set of S/N ratios therefore resembles a discrete dataset. However as the semi-discrete nature of the data is merely a question of precision, we will treat it as continuos data. It should however be noted that the data is highly tied, e.g. $n_{S/N=1.0} = 60,168$, as fig. 4.4.1B clearly illustrates.

4.4.2 Direct signal mapping with standard score normalisation

The initial approach used the entire set of 105, 141 and was set up as a web server called 'PepChipper-1.0'. Figure 4.4.2 and table 4.4.1 gives examples of output from PepChipper-1.0 for VAR2CSA FCR3 sector 5. Full output is available in

Protein	Sector	Start	End	Avg Z	Max Z	Sequence
myprot_FCR3	5	1008	1025	10.212	13.349	CGSARTMKRGYKNDNYEL
myprot_FCR3	5	2676	2686	5.287	6.09	LEGKPIPNPLL

Table 4.4.1: Example of PepChipper-1.0 output for VAR2CSA FCR3 sector 5.

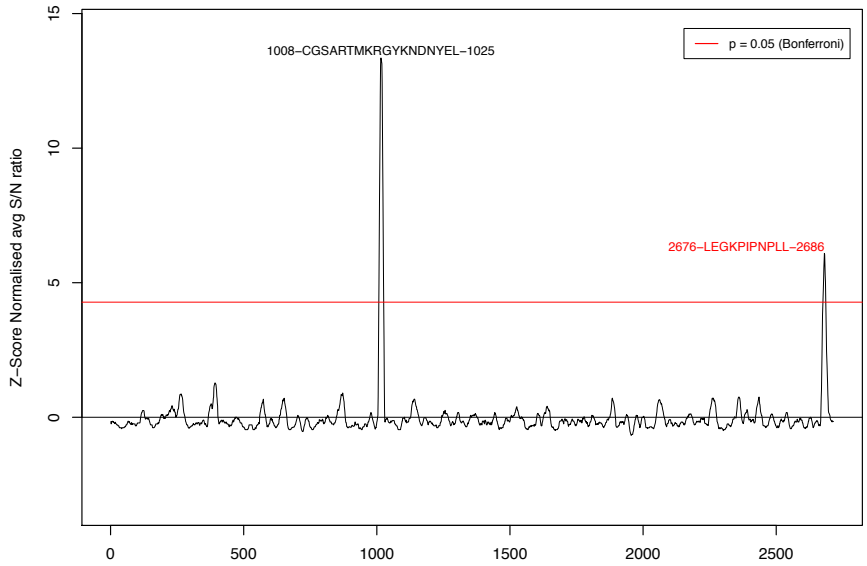


Figure 4.4.2: Plot of scan for linear b-cell epitopes in VAR2CSA FCR3, sector 5. On the x-axis is the VAR2CSA sequence position and on the y-axis the *z-score* normalised average S/N ratio. Significant linear epitopes `1008-CGSARTMKRGYKNDNYEL-1025` and `2676-LEGKPIPNPLL-2686` are identified. The red line signifies Bonferroni-adjusted *z-score* threshold for significance.

supplementary materials.

4.4.3 Identification of epitope linear determinant using a *k-mer* based non-parametric approach

Sector 5 contain 1,087 unique FCR3 15-mers. A total of 11,957 k-mers can be generated of which 2,386 are unique. A *z-score* is computed for each of the 2,386 unique 5-mers. A total of 2,414 5-mers can be positionally mapped to FCR3. Counting each unique 5-mer as a test and setting $\alpha = 0.05$ corresponding to $z_{1-\alpha} = 1.64$, yields an adjusted *z-score* threshold of $z_{1-\alpha/2{,}386} = 4.10$. It should be noted that some of the 5-mers are mapped to more than one position. This is important as the consequence of this is that a given 5-mer may produce a false positive in an irrelevant part of the protein, simply as an artefact of it being present in a true positive in a different part of the analysed protein. Given this frame-work, fig. 4.4.3 depict the mapping of the HDMPa sector 5, 5-mer

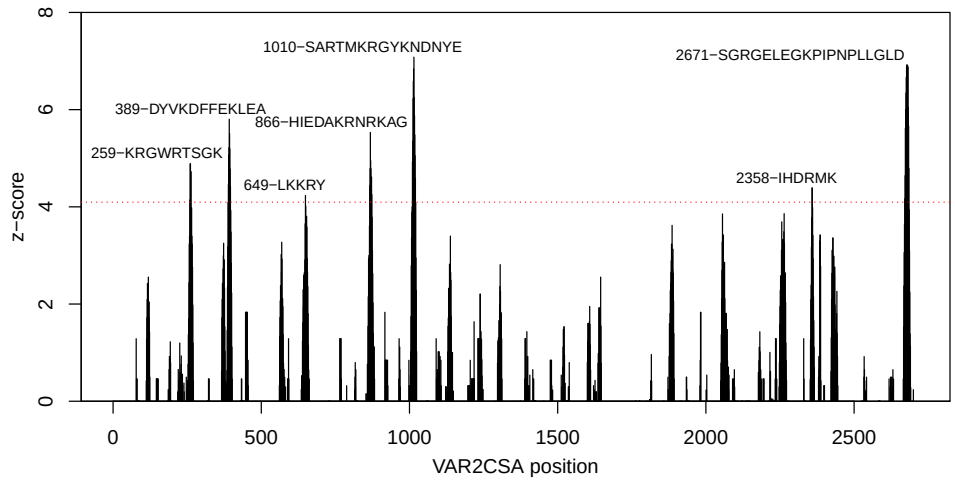


Figure 4.4.3: Plot of results from the analysis of VAR2CSA FCR3 epitope scan data from sector 5 of the high density peptide microarray (HDPMa) chip using the *k-mer* method with $k = 5$. The 5-mers mapped was generated from 1,087 unique FCR3 15-mers on the HDPMa chip, for which $S/N > 1.0$. A total of 11,957 5-mers were generated, of which 2,386 were unique. For each VAR2CSA FCR3 position p_i, $z_{p_i} = z(5mer_{p_i \ldots p_{i+4}}) | z > 0$, i.e. the *z-score* assigned to each position corresponds to the *z-score* of the 5-mer starting at that position, mapping only k-mers for which the *z-score* is larger than 0. Here we apply a one-sided test, since we are only interested in 5-mers with a positive z-score. The red line signifies Bonferroni-adjusted *z-score* threshold for significance. The adjusted *z-score* threshold was obtained by counting each unique 5-mer as a test and then calculating $z_{adj} = z_{1-\alpha/n_{tests}}$, where $\alpha = 0.05$. Continuous stretches of amino acid residues, containing 5-mers for which $z > z_{adj}$ are stated above the respective peak along with the starting position.

population onto FCR3. Table 4.4.2 summarises the identified epitopes.

4.4.4 Evaluation of validity of normal assumption

Fig. 4.4.4 depict quantile-quantile normal plots for the distributions of **A**: The raw signals from all of the 105,141 (3,748 unique) 15-mers and **B**: The *z-scores* derived from analysing 1,156,551 5-mers (3,929 unique).

Protein	Sec	z_{max}	k-mer$_{max}$	Start	End	Sequence
FCR3	5	7.080	KRGYK	1010	1024	SARTMKRGYKNDNYE
FCR3	5	6.920	GKPIP	2671	2689	SGRGELEGKPIPNPLLGLD
FCR3	5	5.803	KDFFE	389	400	DYVKDFFEKLEA
FCR3	5	5.532	EDAKR	866	877	HIEDAKRNRKAG
FCR3	5	4.887	RGWRT	259	267	KRGWRTSGK
FCR3	5	4.389	IHDRM	2358	2363	IHDRMK
FCR3	5	4.233	LKKRY	649	653	LKKRY

Table 4.4.2: FCR3 mapping of 5-mer population. The 5-mers mapped was generated from 1,087 unique FCR3 15-mers from sector 5 on the HDPMa chip. A total of 11,957 5-mers were generated, of which 2,386 are unique.

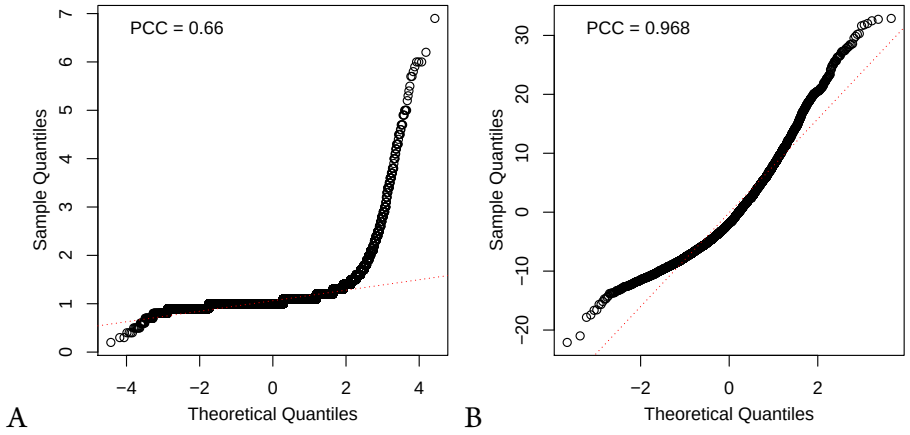


Figure 4.4.4: Quantile-quantile normal plot of **A**: raw signal distribution and **B**: *k-mer* based *z-score* distribution.

4.5 Discussion

The peptide SCGSARTMKRGYKNDNYELCKYC or parts of it, was identified in FCR3 sectors 1,3,4,5,6,7,22,23 and 3D7 sectors 1,3,4,5,6,7,16,18,22,23. As the peptide was observed to elicit a strong response across sectors, it was chosen to as test peptide and used for immunisation and subsequent quantification of immunogenicity and blocking capabilities. These studies were performed by our experimental collaborator at CMP/KU. The peptide was found to be highly immunogenic, but not blocking the VAR2CSA:receptor binding.

Immunodominant epitopes acting as an immunological smoke-screen, diverting response towards non-inhibiting epitopes.

As previously described, epitopes contain a linear determinant ranging from 4 to 7 consecutive amino acid residues. Therefore given a 15-mer with a high signal, it is likely that the signal originates from e.g. 5 consecutive residues, but exactly which 5 residues are responsible for the signal, remains difficult to elude. The second method described attempts to address this by subdividing each 15-mer into k-mers and then assigning the 15-mer signal to each k-mer. This way if a given *k-mer* is consistently associated with a high signal, thusly constituting a 'driver-motif', it should be identifiable on the other hand if a given *k-mer* is found only as a 'passenger-motif', it should also be found in 15-mers with low signal reducing the final rank placement.

The initial method of direct mapping does not take this driver/passenger-motif concept into account, furthermore the risk of quenching information is present. The risk of information quenching arises, because we initially to any non-unique 15-mers assign the mean of the recorded S/N and subsequently each p_i in the reference sequence is assigned the mean of the observed signals. Lastly the normal assumption is not a good fit as fig. 4.4.4A clearly illustrates.

The decision of excluding all 15-mers for which $S/n \leq 1.0$ is substantiated by the fact that z_{max} for the V5 tag `GKPIPNPLLGLDST` is lowered from $z(\texttt{2671-SGRGELE}\underline{\texttt{GKPIP}}\texttt{NPLLGLD-2689}) = 6.920$ to $z(\texttt{2674-GELE}\underline{\texttt{GKPIP}}\texttt{NPLLGLDST-2691}) = 6.341$ if all 15-mers are included in the evaluation. The 5-mer with the highest *z-score* (underlined) does not change. z_{max} for the top-ranking peptide, i.e. the one used for experimental immunisation, is likewise lowered from $z(\texttt{1010-SARTM}\underline{\texttt{KRGYK}}\texttt{NDNYE-1024}) = 7.080$ to $z(\texttt{1008-CGSARTM}\underline{\texttt{KRGYK}}\texttt{NDN-1022}) = 6.399$. Also here the core motif remains constant.

The concept of driver-passenger motifs in quantitative peptide data is illustrated in fig. 4.5.1.

The 3 peptides in fig. 4.5.1 overlaps, such that the C-terminus of peptide 1

No.	Peptide	Signal
1	DNKNQDECQKKLEKV_____	High
2	_____KLEKVFASLTNGYKC_____	High
3	_____NGYKCDKCKSGTSRS	Low

Figure 4.5.1: Driver-passenger motif. The C-terminus of peptide 1 overlaps the N-terminus of peptide 2 and the C-terminus of peptide 2 overlaps the N-terminus of peptide 3 all overlaps are by 5 residues. The driver-motif is highlighted in red and the passenger-motif is highlighted in blue. Peptides 1 and 2 have a high signal because of the driver-motif, despite peptide 2 containing the passenger-motif. Peptide 3 has a low signal, containing only the passenger-motif.

overlaps the N-terminus of peptide 2 and the C-terminus of peptide 2 overlaps the N-terminus of peptide 3, where all overlaps are by 5 residues. The driver-motif is highlighted in red and the passenger motif is highlighted in blue. Peptides 1 and 2 have a high signal because of the driver-motif, despite peptide 2 containing the passenger-motif. Peptide 3 has a low signal, containing only the passenger-motif. The idea of the *k-mer* approach is that dividing the 15-mers into k-mers, we are able to separate the driver-motif from the passenger-motif and thereby the driver-motif will rank better and the passenger-motif worse, within the final test framework.

Part V

Thesis Recapitulation

The offset of the work presented in this thesis was to aid traditional wet-lab experiments, with *in silico* tools for sequence analysis aiming at elucidating the surface variation of VAR2CSA. Moreover: *"The development and application of computational tools aiming at obtaining a better understanding of how VAR2CSA sequence variation affects immunogenicity and capability to induce parasite adhesion blocking antibodies. The ultimate goal being able to produce a vaccine, which can be used to protect pregnant women against Placental Malaria."*

The results presented in this thesis demonstrate how computational approaches can be applied for linear b-cell epitope discovery and how VAR2CSA phenotypes can be correlated with specific immunogenic sequence motifs. Furthermore use-ready *in silico* tools, which hopefully in the future, can bring us closer to a PM vaccine, are described.

Part IIa: By algebraic analysis of the intrinsic properties of the *SigniSite* framework, we have demonstrated how it is possible to reduce the total number of tests performed in a system prior to analysis. The impact of this on the sensitivity of the *SigniSite* framework, i.e. the ability to detect subtle association was however limited. The reason for this is suggested to be due to the non-linearity of the correcting for multiple testing, by dividing the level of significance with the number of tests performed, the function become of type $f(x) = \frac{a}{x}$, which asymptotically approach zero as x increases. Therefore once above a certain number of tests, there is not much difference in the adjusted threshold, e.g. going from 1 to 10 tests, the increase in threshold i of a factor 1.78, whereas going from 10 tests to 1,000 tests, the increase is of a factor 1.16.

*Part IIb:*We analysed 415 MSAs containing from 20 to 43 HLA-A/B sequences, each with an associated binding affinity to the same peptide within each MSA and compared the prediction values from the Shannon entropy and the absolute sum of the positional *SigniSite z-scores* across all 415 MSAs, thereby conducting a meta analysis. In doing so, we found that the two scoring methods were in agreement regarding the ranks of which positions were important in relation to MHC:peptide binding $SCC = 0.996$. We further noted that positions in the pseudo-sequence, were distributed throughout the ranks and thusly

surpassed in rank by positions *not* in the pseudo-sequence. Visual analysis of top-ranking positions mapped onto HLA-A*02:01 furthermore revealed that the top-ranking positions contained amino acid residues, with side-chains protruding into the MHC-I binding groove. Based on these finding, it is proposed to take a closer look at the MHCI pseudo-sequence to see if prediction performance of existing MHCI:peptide predictors can be improved using these top-ranked positions.

Part III: Applying *SigniSite* for the analysis of 70 VAR2CSA-DBL5ε sequences, we found the motif 'TFKNI' to be significantly associated with the birth weight of the newborn. The increased birth weight can be linked to the acquired immunity, as the initial response, will be directed against the VAR2CSA motif, with the greatest capability of inducing a response. Upon following infections, this motif will be recognised and the IE prevented from cyto-adherence. This will give rise to less immunogenic motifs. It has been proposed that the parasite is able of antigenic switching and thusly express not-recognisable motifs. Obtaining a deeper understanding of these mechanisms is crucial in the continuing search for a PM vaccine.

Part IV: Using the HDPMa chip technology for scanning VAR2CSA for linear b-cell epitopes revealed the peptide SCGSARTMKRGYKNDNYELCKYC or parts of it as eliciting a significant signal. Based on this, the peptide was used for immunisation and subsequent quantification of immunigenicity. The peptide was found to be highly immunogenic, but unfortunately the antibodies it induced were not capable of blocking the VAR2CSA:receptor interaktion. This is however not surprising, as it is well known that VAR2CSA contain immunodominant epitopes, against which the humane immune response is diverted, the so called immunological 'smoke-screen'. This ensures that the parasite inside the IE retains the cyte-adherence capabilities. In analysing the HDPMa chip data, two methods were applied. Both methods were in agreement with the findings regarding the before mentioned peptide. However inspecting and comparing epitope plots (see supplementary materials) it seems that the *k-mer* approach is capable of finding epitopes with a more subtle signal. Based on this and the fact that the

HDPMa chip technology has progressed since this study was conducted, it is proposed to run a new chip and apply the *k-mer* method and then subsequently look for epitopes inducing VAR2CSA:receptor blocking antibodies. The fact that immunity toward PM is gradually acquired demonstrate that such epitopes *does* exist. The question regarding this however remains if the linear determinant of the epitope is adequate in inducing antibodies or if the contact-points somehow need to be seen in order to generate a broadly blocking response.

As concluding remark, I would like to stress the importance of continuing the development of methods capable of making genotype-phenotype correlations. Especially in the light of the development of sequencing methods. A sequence without a phenotype really is not that interesting, it is not until we get the phenotype that we are really able to decode the system. However the vast amounts of data being generated poses a challenge in that, as previously mentioned, given enough data, any difference can be significant, no matter how small it may be. Given the increase in the amount of data, we are going from, what has been referred to as *'p-value* hacking' because of small data sets, to a situation, where experimentally based research institutions generate much more data, than they are capable of analysing and the challenge become how to distinguish the 'true' findings from the findings, which arises as a function of the data size, rather than actual biological significance.

Leon Eyrich Jessen
April 2014

References

[1] R N Achur, M Valiyaveettil, A Alkhalil, C F Ockenhouse, and D C Gowda. Characterization of proteoglycans of human placenta and identification of unique chondroitin sulfate proteoglycans of the intervillous spaces that mediate the adherence of Plasmodium falciparum-infected erythrocytes to the placenta. *The Journal of biological chemistry*, 275(51):40344–56, December 2000. ISSN 0021-9258. doi: 10.1074/jbc.M006398200. URL http://www.ncbi.nlm.nih.gov/pubmed/11005814.

[2] Selidji Todagbe Agnandji, Bertrand Lell, José Francisco Fernandes, Béatrice Peggy Abossolo, Barbara Gaelle Nfono Ondo Methogo, Anita Lumeka Kabwende, Ayola Akim Adegnika, Benjamin Mordmüller, Saadou Issifou, Peter Gottfried Kremsner, Jahit Sacarlal, Pedro Aide, Miguel Lanaspa, John J Aponte, Sonia Machevo, Sozinho Acacio, Helder Bulo, Betuel Sigauque, Eusébio Macete, Pedro Alonso, Salim Abdulla, Nahya Salim, Rose Minja, Maxmillian Mpina, Saumu Ahmed, Ali Mohammed Ali, Ali Takadir Mtoro, Ali Said Hamad, Paul Mutani, Marcel Tanner, Halidou Tinto, Umberto D'Alessandro, Hermann Sorgho, Innocent Valea, Biébo Bihoun, Issa Guiraud, Berenger Kaboré, Olivier Sombié, Robert Tinga Guiguemdé, Jean Bosco Ouédraogo, Mary J Hamel, Simon Kariuki, Martina Oneko, Chris Odero, Kephas Otieno, Norbert Awino, Meredith McMorrow, Vincent Muturi-Kioi, Kayla F Laserson, Laurence Slutsker, Walter Otieno, Lucas Otieno, Nekoye Otsyula, Stacey Gondi, Allan Otieno, Victorine Owira, Esther Oguk, George Odongo, Jon Ben Woods, Bernhards Ogutu, Patricia Njuguna, Roma Chilengi, Pauline Akoo, Christine Kerubo, Charity Maingi, Trudie Lang, Ally Olotu, Philip Bejon, Kevin Marsh, Gabriel Mwambingu, Seth Owusu-Agyei, Kwaku Poku Asante, Kingsley Osei-Kwakye, Owusu Boahen, David Dosoo, Isaac Asante, George Adjei, Evans Kwara, Daniel Chandramohan, Brian Greenwood, John Lusingu, Samwel Gesase, Anangisye Malabeja, Omari Abdul, Coline Mahende, Edwin Liheluka,

Lincoln Malle, Martha Lemnge, Thor G Theander, Chris Drakeley, Daniel Ansong, Tsiri Agbenyega, Samuel Adjei, Harry Owusu Boateng, Theresa Rettig, John Bawa, Justice Sylverken, David Sambian, Anima Sarfo, Alex Agyekum, Francis Martinson, Irving Hoffman, Tisungane Mvalo, Portia Kamthunzi, Rutendo Nkomo, Tapiwa Tembo, Gerald Tegha, Mercy Tsidya, Jane Kilembe, Chimwemwe Chawinga, W Ripley Ballou, Joe Cohen, Yolanda Guerra, Erik Jongert, Didier Lapierre, Amanda Leach, Marc Lievens, Opokua Ofori-Anyinam, Aurélie Olivier, Johan Vekemans, Terrell Carter, David Kaslow, Didier Leboulleux, Christian Loucq, Afiya Radford, Barbara Savarese, David Schellenberg, Marla Sillman, and Preeti Vansadia. A phase 3 trial of RTS,S/AS01 malaria vaccine in African infants. *The New England journal of medicine*, 367(24):2284–95, December 2012. ISSN 1533-4406. doi: 10.1056/NEJMoa1208394. URL http://www.ncbi.nlm.nih.gov/pubmed/23136909.

[3] Pedro L Alonso and Marcel Tanner. Public health challenges and prospects for malaria control and elimination. *Nature medicine*, 19(2): 150–5, March 2013. ISSN 1546-170X. doi: 10.1038/nm.3077. URL http://www.ncbi.nlm.nih.gov/pubmed/23389615.

[4] P Armitage, G Berry, and JNS Matthews. *Statistical Methods in Medical Research*. Blackwell Science, Osney Mead, Oxford OX2 0EL, UK, 4th edition, 2002. ISBN 0-632-05257-0.

[5] K Artavanis-Tsakonas, J E Tongren, and E M Riley. The war between the malaria parasite and the immune system: immunity, immunoregulation and immunopathology. *Clinical and experimental immunology*, 133(2): 145–52, August 2003. ISSN 0009-9104. URL http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2249.2003.02174.x/fullhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1808775&tool=pmcentrez&rendertype=abstract.

[6] Alison Deckhut Augustine, B Fenton Hall, Wolfgang W Leitner, Annie X Mo, Tonu M Wali, and Anthony S Fauci. NIAID workshop on immunity to malaria: addressing immunological challenges. *Nature immunology*, 10 (7):673–678, July 2009. ISSN 1529-2916. doi: 10.1038/ni0709-673. URL http://www.ncbi.nlm.nih.gov/pubmed/19536188.

[7] Marion Avril, Megan M Cartwright, Marianne J Hathaway, Mirja Hommel, Salenna R Elliott, Kathryn Williamson, David L Narum,

Patrick E Duffy, Michal Fried, James G Beeson, and Joseph D Smith. Immunization with VAR2CSA-DBL5 recombinant protein elicits broadly cross-reactive antibodies to placental Plasmodium falciparum-infected erythrocytes. *Infection and immunity*, 78(5):2248–56, May 2010. ISSN 1098-5522. doi: 10.1128/IAI.00410-09. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2863527&tool=pmcentrez&rendertype=abstract.

[8] Martin F Bachmann and Gary T Jennings. Vaccine delivery: a matter of size, geometry, kinetics and molecular patterns. *Nature reviews. Immunology*, 10(11):787–96, November 2010. ISSN 1474-1741. doi: 10.1038/nri2868. URL http://www.ncbi.nlm.nih.gov/pubmed/20948547.

[9] Lea Barfod, Nadia L Bernasconi, Madeleine Dahlbäck, David Jarrossay, Pernille Haste Andersen, Ali Salanti, Michael F Ofori, Louise Turner, Mafalda Resende, Morten a Nielsen, Thor G Theander, Federica Sallusto, Antonio Lanzavecchia, and Lars Hviid. Human pregnancy-associated malaria-specific B cells target polymorphic, conformational epitopes in VAR2CSA. *Molecular microbiology*, 63(2):335–47, January 2007. ISSN 0950-382X. doi: 10.1111/j.1365-2958.2006.05503.x. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2779471&tool=pmcentrez&rendertype=abstract.

[10] D J Barlow, M S Edwards, and J M Thornton. Continuous and discontinuous protein antigenic determinants. *Nature*, 322(6081):747–8, 1986. ISSN 0028-0836. doi: 10.1038/322747a0. URL http://www.ncbi.nlm.nih.gov/pubmed/2427953.

[11] D I Baruch, B L Pasloske, H B Singh, X Bi, X C Ma, M Feldman, T F Taraschi, and R J Howard. Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell*, 82(1):77–87, July 1995. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/7541722.

[12] F D Batista and M S Neuberger. B cells extract and present immobilized antigen: implications for affinity discrimination. *The EMBO journal*, 19(4):513–20, March 2000. ISSN 0261-4189. doi: 10.1093/emboj/19.4.513. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=305589&tool=pmcentrez&rendertype=abstract.

[13] Facundo D Batista and Naomi E Harwood. The who, how and where of antigen presentation to B cells. *Nature reviews. Immunology*, 9(1):15–27, January 2009. ISSN 1474-1741. doi: 10.1038/nri2454. URL http://www.ncbi.nlm.nih.gov/pubmed/19079135.

[14] Alan G Baxter and Philip D Hodgkin. Activation rules: the two-signal theories of immune activation. *Nature reviews. Immunology*, 2(6):439–46, June 2002. ISSN 1474-1733. doi: 10.1038/nri823. URL http://www.ncbi.nlm.nih.gov/pubmed/12093010.

[15] Alain Beck, Thierry Wurch, Christian Bailly, and Nathalie Corvaia. Strategies and challenges for the next generation of therapeutic antibodies. *Nature reviews. Immunology*, 10(5):345–52, May 2010. ISSN 1474-1741. doi: 10.1038/nri2747. URL http://www.ncbi.nlm.nih.gov/pubmed/20414207.

[16] J G Beeson, J C Reeder, S J Rogerson, and G V Brown. Parasite adhesion and immune evasion in placental malaria. *Trends in parasitology*, 17(7): 331–7, July 2001. ISSN 1471-4922. URL http://www.ncbi.nlm.nih.gov/pubmed/11423376.

[17] James G Beeson, Jo-Anne Chan, and Freya J I Fowkes. PfEMP1 as a target of human immunity and a vaccine candidate against malaria. *Expert review of vaccines*, 12(2):105–8, February 2013. ISSN 1744-8395. doi: 10.1586/erv.12.144. URL http://www.ncbi.nlm.nih.gov/pubmed/23414401.

[18] Elaine Bell. Who was Edward Jenner? *Nature Reviews Immunology*, 3 (February):2003, 2003. doi: 10.1038/nri1016. URL http://www.nature.com/nri/journal/v3/n2/full/nri1016.html.

[19] Graham A Bentley and Benoît Gamain. How does Plasmodium falciparum stick to CSA? Let's see in the crystal. *Nature structural & molecular biology*, 15(9):895–7, September 2008. ISSN 1545-9993. doi: 10.1038/nsmb0908-895. URL http://www.nature.com/nsmb/journal/v15/n9/abs/nsmb0908-895.htmlhttp://www.ncbi.nlm.nih.gov/pubmed/18769465.

[20] Alain Bernard, Laurence Lamy And, and Isabelle Alberti. The two-signal model of T-cell activation after 30 years. *Transplantation*, 73(1 Suppl):

S31–5, January 2002. ISSN 0041-1337. URL http://www.ncbi.nlm.nih.gov/pubmed/11810059.

[21] Gwladys I Bertin, Thomas Lavstsen, François Guillonneau, Justin Doritchamou, Christian W Wang, Jakob S Jespersen, Sem Ezimegnon, Nadine Fievet, Maroufou J Alao, Francis Lalya, Achille Massougbodji, Nicaise Tuikue Ndam, Thor G Theander, and Philippe Deloron. Expression of the domain cassette 8 Plasmodium falciparum erythrocyte membrane protein 1 is associated with cerebral malaria in Benin. *PloS one*, 8(7):e68368, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0068368. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3726661&tool=pmcentrez&rendertype=abstract.

[22] Francisco a Bonilla and Hans C Oettgen. Adaptive immunity. *The Journal of allergy and clinical immunology*, 125(2 Suppl 2):S33–40, March 2010. ISSN 1097-6825. doi: 10.1016/j.jaci.2009.09.017. URL http://www.ncbi.nlm.nih.gov/pubmed/20061006.

[23] Alan Brown, Louise Turner, Stig Christoffersen, Katrina a Andrews, Tadge Szestak, Yuguang Zhao, Sine Larsen, Alister G Craig, and Matthew K Higgins. Molecular architecture of a complex between an adhesion protein from the malaria parasite and intracellular adhesion molecule 1. *The Journal of biological chemistry*, 288(8):5992–6003, February 2013. ISSN 1083-351X. doi: 10.1074/jbc.M112.416347. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3581401&tool=pmcentrez&rendertype=abstract.

[24] P a Buffet, B Gamain, C Scheidig, D Baruch, J D Smith, R Hernandez-Rivas, B Pouvelle, S Oishi, N Fujii, T Fusai, D Parzy, L H Miller, J Gysin, and A Scherf. Plasmodium falciparum domain mediating adhesion to chondroitin sulfate A: a receptor for human placental infection. *Proceedings of the National Academy of Sciences of the United States of America*, 96(22):12743–8, October 1999. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=23079&tool=pmcentrez&rendertype=abstract.

[25] Sø ren Buus, Johan Rockberg, Björn Forsström, Peter Nilsson, Mathias Uhlen, and Claus Schafer-Nielsen. High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays. *Molecular*

& cellular proteomics : MCP, 11(12):1790–800, December 2012. ISSN 1535-9484. doi: 10.1074/mcp.M112.020800. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3518105&tool=pmcentrez&rendertype=abstract.

[26] P a Carvalho, M Diez-Silva, H Chen, M Dao, and S Suresh. Cytoadherence of erythrocytes invaded by Plasmodium falciparum: quantitative contact-probing of a human malaria receptor. *Acta biomaterialia*, 9(5):6349–59, May 2013. ISSN 1878-7568. doi: 10.1016/j.actbio.2013.01.019. URL http://www.ncbi.nlm.nih.gov/pubmed/23376131.

[27] Sofia Casares and Thomas L Richie. Immune evasion by malaria parasites: a challenge for vaccine development. *Current opinion in immunology*, 21 (3):321–30, June 2009. ISSN 1879-0372. doi: 10.1016/j.coi.2009.05.015. URL http://www.ncbi.nlm.nih.gov/pubmed/19493666.

[28] Andrew C Chan and Paul J Carter. Therapeutic antibodies for autoimmunity and inflammation. *Nature reviews. Immunology*, 10(5): 301–16, May 2010. ISSN 1474-1741. doi: 10.1038/nri2761. URL http://www.ncbi.nlm.nih.gov/pubmed/20414204.

[29] Jo-anne Chan, Katherine B Howell, Linda Reiling, Ricardo Ataide, Claire L Mackintosh, Freya J I Fowkes, Michaela Petter, Joanne M Chesson, Christine Langer, George M Warimwe, Michael F Duffy, Stephen J Rogerson, Peter C Bull, Alan F Cowman, Kevin Marsh, and James G Beeson. Targets of antibodies against Plasmodium falciparum-infected erythrocytes in malaria immunity. *The Journal of clinical investigation*, 122(9):3227–38, September 2012. ISSN 1558-8238. doi: 10.1172/JCI62182. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3428085&tool=pmcentrez&rendertype=abstract.

[30] David D Chaplin. Overview of the human immune response. *The Journal of allergy and clinical immunology*, 117(2 Suppl Mini-Primer):S430–5, March 2006. ISSN 0091-6749. doi: 10.1016/j.jaci.2005.09.034. URL http://www.ncbi.nlm.nih.gov/pubmed/16455341.

[31] David D Chaplin. Overview of the immune response. *The Journal of allergy and clinical immunology*, 125(2 Suppl 2):S3–23, March 2010. ISSN

1097-6825. doi: 10.1016/j.jaci.2009.12.980. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2923430&tool=pmcentrez&rendertype=abstract.

[32] Monika Chugh, Vidhya Sundararaman, Saravanan Kumar, Vanga S Reddy, Waseem A Siddiqui, Kenneth D Stuart, and Pawan Malhotra. Protein complex directs hemoglobin-to-hemozoin formation in Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America*, 110(14):5392–7, April 2013. ISSN 1091-6490. doi: 10.1073/pnas.1218412110. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3619337&tool=pmcentrez&rendertype=abstract.

[33] Thomas M. Clausen, Stig Christoffersen, Madeleine Dahlbäck, Annette Eva Langkilde, Kamilla E. Jensen, Mafalda Resende, Mette ØAgerbæ k, Daniel Andersen, Besim Berisha, Sisse B. Ditlev, Vera V. Pinto, Morten a. Nielsen, Thor G. Theander, Sine Larsen, and Ali Salanti. Structural and functional insight into how the Plasmodium falciparum VAR2CSA protein mediates binding to chondroitin sulfate A in placental malaria. *The Journal of biological chemistry*, 287(28): 23332–45, July 2012. ISSN 1083-351X. doi: 10.1074/jbc.M112.348839. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3390611&tool=pmcentrez&rendertype=abstract.

[34] Irun R Cohen. Real and artificial immune systems: computing the state of the body. *Nature reviews. Immunology*, 7(7):569–74, July 2007. ISSN 1474-1733. doi: 10.1038/nri2102. URL http://www.ncbi.nlm.nih.gov/pubmed/17558422.

[35] Max D Cooper and Brantley R Herrin. How did our complex immune system evolve? *Nature reviews. Immunology*, 10(1):2–3, January 2010. ISSN 1474-1741. doi: 10.1038/nri2686. URL http://www.ncbi.nlm.nih.gov/pubmed/20039476.

[36] a Corthay. A three-cell model for activation of naive T helper cells. *Scandinavian Journal of Immunology*, 64(0300-9475 (Print) LA - eng PT - Comment PT - Journal Article PT - Research Support, Non-U.S. Gov't PT - Review RN - 207137-56-2 (Interleukin-4) RN - 82115-62-6 (Interferon Type II) SB - IM):93–96, August 2006. ISSN 0300-9475. doi:

10.1111/j.1365-3083.2006.01782.x. URL http://www.ncbi.nlm.nih.gov/pubmed/16867153.

[37] Alan F Cowman and Brendan S Crabb. Invasion of red blood cells by malaria parasites. *Cell*, 124(4):755–66, February 2006. ISSN 0092-8674. doi: 10.1016/j.cell.2006.02.006. URL http://www.ncbi.nlm.nih.gov/pubmed/16497586.

[38] Alan F Cowman, Drew Berry, and Jake Baum. The cellular and molecular basis for malaria parasite invasion of the human red blood cell. *The Journal of cell biology*, 198(6):961–71, September 2012. ISSN 1540-8140. doi: 10.1083/jcb.201206112. URL http://www.ncbi.nlm.nih.gov/pubmed/22986493$\delimiter"026E30F$nhttp://jcb.rupress.org/content/198/6/961.shorthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3444787&tool=pmcentrez&rendertype=abstract.

[39] Madeleine Dahlbäck, Thomas S Rask, Pernille H Andersen, Morten a Nielsen, Nicaise T Ndam, Mafalda Resende, Louise Turner, Philippe Deloron, Lars Hviid, Ole Lund, Anders Gorm Pedersen, Thor G Theander, and Ali Salanti. Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in P. falciparum placental sequestration. *PLoS pathogens*, 2(11):e124, November 2006. ISSN 1553-7374. doi: 10.1371/journal.ppat.0020124. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1636682&tool=pmcentrez&rendertype=abstract.

[40] Madeleine Dahlbäck, Lars M Jø rgensen, Morten a Nielsen, Thomas M Clausen, Sisse B Ditlev, Mafalda Resende, Vera V Pinto, David E Arnot, Thor G Theander, and Ali Salanti. The chondroitin sulfate A-binding site of the VAR2CSA protein involves multiple N-terminal domains. *The Journal of biological chemistry*, 286(18):15908–17, May 2011. ISSN 1083-351X. doi: 10.1074/jbc.M110.191510. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3091200&tool=pmcentrez&rendertype=abstract.

[41] Sisse B Ditlev, Raluca Florea, Morten a Nielsen, Thor G Theander, Stefan Magez, Philippe Boeuf, and Ali Salanti. Utilizing Nanobody Technology to Target Non-Immunodominant Domains of VAR2CSA. *PloS one*, 9(1): e84981, January 2014. ISSN 1932-6203. doi:

10.1371/journal.pone.0084981. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3897377&tool=pmcentrez&rendertype=abstract.

[42] Denise L Doolan, Carlota Dobaño, and J Kevin Baird. Acquired immunity to malaria. *Clinical microbiology reviews*, 22(1):13–36, Table of Contents, January 2009. ISSN 1098-6618. doi: 10.1128/CMR.00025-08. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2620631&tool=pmcentrez&rendertype=abstract.

[43] Chris Drakeley, Colin Sutherland, J Teun Bousema, Robert W Sauerwein, and Geoffrey a T Targett. The epidemiology of Plasmodium falciparum gametocytes: weapons of mass dispersion. *Trends in parasitology*, 22(9): 424–30, September 2006. ISSN 1471-4922. doi: 10.1016/j.pt.2006.07.001. URL http://www.ncbi.nlm.nih.gov/pubmed/16846756.

[44] S Dudoit, Y H Yang, M J Callow, and T P Speed. Statistical methods for identifying differentially expressed genes in replicated c{DNA} microarray experiments. *Stat. Sinica*, 12(1):111–139, 2002.

[45] Patrick E Duffy and Robert S. Desowitz. Pregnancy malaria throughout history: dangerous labors. In Patrick E Duffy and Michael Fried, editors, *Malaria in pregnancy: deadly parasite, susceptible host*, pages 1–25. Informa Healthcare, Zug, Canton of Zug, Switzerland, 2001. ISBN 0415272181.

[46] a a Escalante and F J Ayala. Phylogeny of the malarial genus Plasmodium, derived from rRNA gene sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 91(24):11373–7, November 1994. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=45233&tool=pmcentrez&rendertype=abstract.

[47] Rachel E Farrow, Judith Green, Zoe Katsimitsoulia, William R Taylor, Anthony a Holder, and Justin E Molloy. The mechanism of erythrocyte invasion by the malarial parasite, Plasmodium falciparum. *Seminars in cell & developmental biology*, 22(9):953–60, December 2011. ISSN 1096-3634. doi: 10.1016/j.semcdb.2011.09.022. URL http://www.ncbi.nlm.nih.gov/pubmed/22001249.

[48] Michael Föller, Diwakar Bobbala, Saisudha Koka, Stephan M Huber, Erich Gulbins, and Florian Lang. Suicide for survival–death of infected

erythrocytes as a host mechanism to survive malaria. *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology*, 24(3-4):133–40, January 2009. ISSN 1421-9778. doi: 10.1159/000233238. URL http://www.ncbi.nlm.nih.gov/pubmed/19710527.

[49] Lander Foquet, Cornelus C Hermsen, Geert-Jan van Gemert, Eva Van Braeckel, Karin E Weening, Robert Sauerwein, Philip Meuleman, and Geert Leroux-Roels. Vaccine-induced monoclonal antibodies targeting circumsporozoite protein prevent Plasmodium falciparum infection. *The Journal of clinical investigation*, 124(1), December 2013. ISSN 1558-8238. doi: 10.1172/JCI70349. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3871238&tool=pmcentrez&rendertype=abstract.

[50] M Fried and P E Duffy. Adherence of Plasmodium falciparum to chondroitin sulfate A in the human placenta. *Science (New York, N.Y.)*, 272 (5267):1502–4, June 1996. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/8633247.

[51] M Fried, F Nosten, A Brockman, B J Brabin, and P E Duffy. Maternal antibodies block malaria. *Nature*, 395(6705):851–2, October 1998. ISSN 0028-0836. doi: 10.1038/27570. URL http://www.ncbi.nlm.nih.gov/pubmed/9804416.

[52] Carlo a J M Gaillard and Raymond M Schiffelers. Red blood cell: barometer of cardiovascular health? *Cardiovascular research*, 98(1):3–4, April 2013. ISSN 1755-3245. doi: 10.1093/cvr/cvt041. URL http://www.ncbi.nlm.nih.gov/pubmed/23428359.

[53] Richard L Gallo and Lora V Hooper. Epithelial antimicrobial defence of the skin and intestine. *Nature reviews. Immunology*, 12(7):503–16, July 2012. ISSN 1474-1741. doi: 10.1038/nri3228. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3563335&tool=pmcentrez&rendertype=abstract.

[54] Stéphane Gangnard, Cyril Badaut, Stéphanie Ramboarina, Bruno Baron, Tarik Ramdani, Benoît Gamain, Philippe Deloron, Anita Lewit-Bentley, and Graham a Bentley. Structural and immunological correlations between the variable blocks of the VAR2CSA domain DBL6ε from two

Plasmodium falciparum parasite lines. *Journal of molecular biology*, 425 (10):1697–711, May 2013. ISSN 1089-8638. doi: 10.1016/j.jmb.2013.02.014. URL http://www.ncbi.nlm.nih.gov/pubmed/23429057.

[55] Tomas Ganz. Defensins: antimicrobial peptides of innate immunity. *Nature reviews. Immunology*, 3(9):710–20, September 2003. ISSN 1474-1733. doi: 10.1038/nri1180. URL http://www.ncbi.nlm.nih.gov/pubmed/12949495.

[56] Tomas Ganz. Macrophages and systemic iron homeostasis. *Journal of innate immunity*, 4(5-6):446–53, January 2012. ISSN 1662-8128. doi: 10.1159/000336423. URL http://www.ncbi.nlm.nih.gov/pubmed/22441209.

[57] Maria Garcia-Boronat, Carmen M Diez-Rivero, Ellis L Reinherz, and Pedro a Reche. PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. *Nucleic acids research*, 36 (Web Server issue):W35–41, July 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn211. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447719&tool=pmcentrez&rendertype=abstract.

[58] Mehrdad Ghashghaeinia, Judith C a Cluitmans, Ahmed Akel, Peter Dreischer, Mahmoud Toulany, Martin Köberle, Yuliya Skabytska, Mohammad Saki, Tilo Biedermann, Michael Duszenko, Florian Lang, Thomas Wieder, and Giel J C G M Bosman. The impact of erythrocyte age on eryptosis. *British journal of haematology*, 157(5):606–14, June 2012. ISSN 1365-2141. doi: 10.1111/j.1365-2141.2012.09100.x. URL http://www.ncbi.nlm.nih.gov/pubmed/22429222.

[59] Paul R Gilson and Brendan S Crabb. Morphology and kinetics of the three distinct phases of red blood cell invasion by Plasmodium falciparum merozoites. *International journal for parasitology*, 39(1):91–6, January 2009. ISSN 1879-0135. doi: 10.1016/j.ijpara.2008.09.007. URL http://www.ncbi.nlm.nih.gov/pubmed/18952091.

[60] Jacob Glanville, Wenwu Zhai, Jan Berka, Dilduz Telman, Gabriella Huerta, Gautam R Mehta, Irene Ni, Li Mei, Purnima D Sundar, Giles M R Day, David Cox, Arvind Rajpal, and Jaume Pons. Precise determination of

the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences of the United States of America*, 106(48):20216–21, December 2009. ISSN 1091-6490. doi: 10.1073/pnas.0909775106. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2787155&tool=pmcentrez&rendertype=abstract.

[61] Sédami Gnidehou, Leon Jessen, Stéphane Gangnard, Caroline Ermont, Choukri Triqui, Mickael Quiviger, Juliette Guitard, Ole Lund, Philippe Deloron, and Nicaise Tuikue Ndam. Insight into antigenic diversity of VAR2CSA-DBL5ε domain from multiple Plasmodium falciparum placental isolates. *PloS one*, 5(10), January 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0013105. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2948511&tool=pmcentrez&rendertype=abstract.

[62] M F Good. Towards a blood-stage vaccine for malaria: are we following all the leads? *Nature reviews. Immunology*, 1(2):117–25, November 2001. ISSN 1474-1733. doi: 10.1038/35100540. URL http://www.ncbi.nlm.nih.gov/pubmed/11905819.

[63] Kasturi Haldar, Sophien Kamoun, N Luisa Hiller, Souvik Bhattacharje, and Christiaan van Ooij. Common infection strategies of pathogenic eukaryotes. *Nature reviews. Microbiology*, 4(12):922–31, December 2006. ISSN 1740-1534. doi: 10.1038/nrmicro1549. URL http://www.ncbi.nlm.nih.gov/pubmed/17088934.

[64] Lajla Bruntse Hansen, Soren Buus, and Claus Schafer-Nielsen. Identification and mapping of linear antibody epitopes in human serum albumin using high-density Peptide arrays. *PloS one*, 8(7):e68902, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0068902. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3720873&tool=pmcentrez&rendertype=abstract.

[65] Jürgen Harder, Stefanie Dressel, Maike Wittersheim, Jesko Cordes, Ulf Meyer-Hoffert, Ulrich Mrowietz, Regina Fölster-Holst, Ehrhard Proksch, Jens-Michael Schröder, Thomas Schwarz, and Regine Gläser. Enhanced expression and secretion of antimicrobial peptides in atopic dermatitis and after superficial skin injury. *The Journal of investigative dermatology*,

130(5):1355–64, May 2010. ISSN 1523-1747. doi: 10.1038/jid.2009.432. URL http://www.ncbi.nlm.nih.gov/pubmed/20107483.

[66] Pernille Haste Andersen, Morten Nielsen, and Ole Lund. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein science : a publication of the Protein Society*, 15(11):2558–67, November 2006. ISSN 0961-8368. doi: 10.1110/ps.062405906. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2242418&tool=pmcentrez&rendertype=abstract.

[67] Ernst Hempelmann and Kristine Krafts. Bad air, amulets and mosquitoes: 2,000 years of changing perspectives on malaria. *Malaria journal*, 12(1): 232, July 2013. ISSN 1475-2875. doi: 10.1186/1475-2875-12-232. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3723432&tool=pmcentrez&rendertype=abstract.

[68] Matthew K Higgins. The structure of a chondroitin sulfate-binding domain important in placental malaria. *The Journal of biological chemistry*, 283(32):21842–6, August 2008. ISSN 0021-9258. doi: 10.1074/jbc.C800086200. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2494935&tool=pmcentrez&rendertype=abstract.

[69] Matthew K Higgins and Mark Carrington. Sequence variation and structural conservation allows development of novel function and immune evasion in parasite surface protein families. *Protein science : a publication of the Protein Society*, page Epub ahead of print, January 2014. ISSN 1469-896X. doi: 10.1002/pro.2428. URL http://www.ncbi.nlm.nih.gov/pubmed/24442723.

[70] Adrian V S Hill. Pre-erythrocytic malaria vaccines: towards greater efficacy. *Nature reviews. Immunology*, 6(1):21–32, January 2006. ISSN 1474-1733. doi: 10.1038/nri1746. URL http://www.ncbi.nlm.nih.gov/pubmed/16493425.

[71] R C Hillig, P G Coulie, V Stroobant, W Saenger, A Ziegler, and M Hülsmeyer. High-resolution structure of HLA-A*0201 in complex with a tumour-specific antigenic peptide encoded by the MAGE-A4 gene. *Journal of molecular biology*, 310(5):1167–76, July 2001. ISSN 0022-2836. doi: 10.1006/jmbi.2001.4816. URL http://www.ncbi.nlm.nih.gov/pubmed/11502003.

[72] Hajime Hisaeda, Koji Yasutomo, and Kunisuke Himeno. Malaria: immune evasion by parasites. *The international journal of biochemistry & cell biology*, 37(4):700–6, April 2005. ISSN 1357-2725. doi: 10.1016/j.biocel.2004.10.009. URL http://www.ncbi.nlm.nih.gov/pubmed/15694829.

[73] M Ho and Nicholas J White. Molecular mechanisms of cytoadherence in malaria. *The American journal of physiology*, 276(6 Pt 1):C1231–42, June 1999. ISSN 0002-9513. URL http://www.ncbi.nlm.nih.gov/pubmed/10362584.

[74] Stephen L Hoffman, Peter F Billingsley, Eric James, Adam Richman, Mark Loyevsky, Tao Li, Sumana Chakravarty, Anusha Gunasekera, Rana Chattopadhyay, Minglin Li, Richard Stafford, Adriana Ahumada, Judith E Epstein, Martha Sedegah, Sharina Reyes, Thomas L Richie, Kirsten E Lyke, Robert Edelman, Matthew B Laurens, Christopher V Plowe, and B Kim Lee Sim. Development of a metabolically active, non-replicating sporozoite vaccine to prevent Plasmodium falciparum malaria. *Human vaccines*, 6(1):97–106, January 2010. ISSN 1554-8619. URL http://www.ncbi.nlm.nih.gov/pubmed/19946222.

[75] L Hviid and A Salanti. VAR2CSA and protective immunity against pregnancy-associated Plasmodium falciparum malaria. *Parasitology*, 134 (Pt 13):1871–6, January 2007. ISSN 0031-1820. doi: 10.1017/S0031182007000121. URL http://www.ncbi.nlm.nih.gov/pubmed/17958922.

[76] Lars Hviid. The case for PfEMP1-based vaccines to protect pregnant women against Plasmodium falciparum malaria. *Expert review of vaccines*, 10(10):1405–14, October 2011. ISSN 1744-8395. doi: 10.1586/erv.11.113. URL http://www.ncbi.nlm.nih.gov/pubmed/21988306.

[77] Darrell J Irvine, Melody a Swartz, and Gregory L Szeto. Engineering synthetic vaccines using cues from natural immunity. *Nature materials*, 12 (11):978–90, November 2013. ISSN 1476-1122. doi: 10.1038/nmat3775. URL http://www.ncbi.nlm.nih.gov/pubmed/24150416.

[78] Leon Eyrich Jessen, Ilka Hoof, Ole Lund, and Morten Nielsen. SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments. *Nucleic acids research*, 41(Web Server

issue):W286–91, July 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt497. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3692133&tool=pmcentrez&rendertype=abstract.

[79] Peng Ji, Maki Murata-Hori, and Harvey F Lodish. Formation of mammalian erythrocytes: chromatin condensation and enucleation. *Trends in cell biology*, 21(7):409–15, July 2011. ISSN 1879-3088. doi: 10.1016/j.tcb.2011.04.003. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3134284&tool=pmcentrez&rendertype=abstract.

[80] Louise Joergensen, Dominique C Bengtsson, Anja Bengtsson, Elena Ronander, Sanne S Berger, Louise Turner, Michael B Dalgaard, Gerald K K Cham, Michala E Victor, Thomas Lavstsen, Thor G Theander, David E Arnot, and Anja T R Jensen. Surface co-expression of two different PfEMP1 antigens on single plasmodium falciparum-infected erythrocytes facilitates binding to ICAM1 and PECAM1. *PLoS pathogens*, 6(9):e1001083, January 2010. ISSN 1553-7374. doi: 10.1371/journal.ppat.1001083. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2932717&tool=pmcentrez&rendertype=abstract.

[81] Stefan H E Kaufmann. The contribution of immunology to the rational design of novel antibacterial vaccines. *Nature reviews. Microbiology*, 5(7): 491–504, July 2007. ISSN 1740-1534. doi: 10.1038/nrmicro1688. URL http://www.ncbi.nlm.nih.gov/pubmed/17558425.

[82] Pongsak Khunrae, Judith M D Philip, Duncan R Bull, and Matthew K Higgins. Structural comparison of two CSPG-binding DBL domains from the VAR2CSA protein important in malaria during pregnancy. *Journal of molecular biology*, 393(1):202–13, October 2009. ISSN 1089-8638. doi: 10.1016/j.jmb.2009.08.027. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3778748&tool=pmcentrez&rendertype=abstract.

[83] Pongsak Khunrae, Madeleine Dahlbäck, Morten a Nielsen, Gorm Andersen, Sisse B Ditlev, Mafalda Resende, Vera V Pinto, Thor G Theander, Matthew K Higgins, and Ali Salanti. Full-length recombinant Plasmodium falciparum VAR2CSA binds specifically to CSPG and induces potent parasite adhesion-blocking antibodies. *Journal of molecular*

biology, 397(3):826–34, April 2010. ISSN 1089-8638. doi: 10.1016/j.jmb.2010.01.040. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3715698&tool=pmcentrez&rendertype=abstract.

[84] Thomas J. Kindt, Richard A. Goldsby, and Barbara A. Osborne. *Immunology*. W. H. Freeman and Company, New York, 6th edition, 2007. ISBN 1429202114.

[85] Laura a Kirkman and Kirk W Deitsch. Antigenic variation and the generation of diversity in malaria parasites. *Current opinion in microbiology*, 15(4):456–62, August 2012. ISSN 1879-0364. doi: 10.1016/j.mib.2012.03.003. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3399988&tool=pmcentrez&rendertype=abstract.

[86] Susan M Kraemer and Joseph D Smith. A family affair: var genes, PfEMP1 binding, and malaria disease. *Current opinion in microbiology*, 9 (4):374–80, August 2006. ISSN 1369-5274. doi: 10.1016/j.mib.2006.06.006. URL http://www.ncbi.nlm.nih.gov/pubmed/16814594.

[87] Jens Vindahl Kringelum, Morten Nielsen, Søren Berg Padkjæ r, and Ole Lund. Structural analysis of B-cell epitopes in antibody:protein complexes. *Molecular immunology*, 53(1-2):24–34, January 2013. ISSN 1872-9142. doi: 10.1016/j.molimm.2012.06.001. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3461403&tool=pmcentrez&rendertype=abstract.

[88] S Kullback and RA Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 1951. URL http://www.jstor.org/stable/10.2307/2236703.

[89] Florian Lang, Elisabeth Lang, and Michael Föller. Physiology and pathophysiology of eryptosis. *Transfusion medicine and hemotherapy : offizielles Organ der Deutschen Gesellschaft fur Transfusionsmedizin und Immunhamatologie*, 39(5):308–14, October 2012. ISSN 1660-3796. doi: 10.1159/000342534. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3678267&tool=pmcentrez&rendertype=abstract.

[90] Jean Langhorne, Francis M Ndungu, Anne-Marit Sponaas, and Kevin Marsh. Immunity to malaria: more questions than answers. *Nature immunology*, 9(7):725–32, July 2008. ISSN 1529-2916. doi: 10.1038/ni.f.205. URL http://www.ncbi.nlm.nih.gov/pubmed/18563083.

[91] Lewis L Lanier. NK cell recognition. *Annual review of immunology*, 23: 225–74, January 2005. ISSN 0732-0582. doi: 10.1146/annurev.immunol.23.021704.115526. URL http://www.ncbi.nlm.nih.gov/pubmed/15771571.

[92] Lewis L. Lanier. Shades of grey — the blurring view of innate and adaptive immunity. *Nature Reviews Immunology*, 13(2):73–74, January 2013. ISSN 1474-1733. doi: 10.1038/nri3389. URL http://www.nature.com/doifinder/10.1038/nri3389.

[93] Lewis L Lanier and Joseph C Sun. Do the terms innate and adaptive immunity create conceptual barriers? *Nature reviews. Immunology*, 9(5): 302–3, May 2009. ISSN 1474-1741. doi: 10.1038/nri2547. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2844347&tool=pmcentrez&rendertype=abstract.

[94] Stewart H Lecker, Alfred L Goldberg, and William E Mitch. Protein degradation by the ubiquitin-proteasome pathway in normal and disease states. *Journal of the American Society of Nephrology : JASN*, 17(7): 1807–19, July 2006. ISSN 1046-6673. doi: 10.1681/ASN.2006010083. URL http://www.ncbi.nlm.nih.gov/pubmed/16738015.

[95] Sang-won Lee, Philip F Markham, Mauricio J C Coppo, Alistair R Legione, John F Markham, Amir H Noormohammadi, Glenn F Browning, Nino Ficorilli, Carol A Hartley, and Joanne M Devlin. Attenuated vaccines can recombine to form virulent field viruses. *Science (New York, N.Y.)*, 337(6091):188, July 2012. ISSN 1095-9203. doi: 10.1126/science.1217134. URL http://www.ncbi.nlm.nih.gov/pubmed/22798607.

[96] Gary W Litman, Jonathan P Rast, and Sebastian D Fugmann. The origins of vertebrate adaptive immunity. *Nature reviews. Immunology*, 10(8): 543–53, August 2010. ISSN 1474-1741. doi: 10.1038/nri2807. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2919748&tool=pmcentrez&rendertype=abstract.

[97] F B Livingstone. Malaria and human polymorphisms. *Annual review of genetics*, 5(3):33–64, January 1971. ISSN 0066-4197. doi: 10.1146/annurev.ge.05.120171.000341. URL http://www.annualreviews.org/doi/pdf/10.1146/annurev.ge.05.120171.000341http://www.ncbi.nlm.nih.gov/pubmed/2680886http://www.ncbi.nlm.nih.gov/pubmed/16097650.

[98] Thomas C. Luke and Stephen L Hoffman. Rationale and plans for developing a non-replicating, metabolically active, radiation-attenuated Plasmodium falciparum sporozoite vaccine. *The Journal of experimental biology*, 206(Pt 21):3803–8, November 2003. ISSN 0022-0949. doi: 10.1242/jeb.00644. URL http://jeb.biologists.org/cgi/doi/10.1242/jeb.00644http://www.ncbi.nlm.nih.gov/pubmed/14506215.

[99] Ole Lund, Morten Nielsen, Claus Lundegaard, Can Kesmir, and Sø ren Brunak. *Immunological Bioinformatics*. The MIT Press, Cambridge, Massachusetts, London, England, 1st edition, 2005. ISBN 0262122804.

[100] Claus Lundegaard, Kasper Lamberth, Mikkel Harndahl, Sø ren Buus, Ole Lund, and Morten Nielsen. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic acids research*, 36(Web Server issue): W509–12, July 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn202. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447772&tool=pmcentrez&rendertype=abstract.

[101] W MacARTHUR. A brief story of English malaria. *Postgraduate medical journal*, 22(249):198–200, July 1946. ISSN 0032-5473. doi: 10.1136/pgmj.22.249.198. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2478349&tool=pmcentrez&rendertype=abstract.

[102] Alexander G Maier, Brian M Cooke, Alan F Cowman, and Leann Tilley. Malaria parasite proteins that remodel the host erythrocyte. *Nature reviews. Microbiology*, 7(5):341–54, May 2009. ISSN 1740-1534. doi: 10.1038/nrmicro2110. URL http://www.ncbi.nlm.nih.gov/pubmed/19369950.

[103] K Marsh and S Kinyanjui. Immune effector mechanisms in malaria. *Parasite immunology*, 28(1-2):51–60, 2006. ISSN 0141-9838. doi:

10.1111/j.1365-3024.2006.00808.x. URL http://www.ncbi.nlm.nih.gov/pubmed/16438676.

[104] Michael McHeyzer-Williams, Shinji Okitsu, Nathaniel Wang, and Louise McHeyzer-Williams. Molecular programming of B cell memory. *Nature reviews. Immunology*, 12(1):24–34, January 2012. ISSN 1474-1741. doi: 10.1038/nri3128. URL http://www.ncbi.nlm.nih.gov/pubmed/22158414.

[105] Louis H Miller, Dror I Baruch, Kevin Marsh, and Ogobara K Doumbo. The pathogenic basis of malaria. *Nature*, 415(6872):673–9, February 2002. ISSN 0028-0836. doi: 10.1038/415673a. URL http://www.ncbi.nlm.nih.gov/pubmed/11832955.

[106] Frank P Mockenhaupt, George Bedu-Addo, Christiane von Gaertner, Renate Boyé, Katrin Fricke, Iris Hannibal, Filiz Karakaya, Marieke Schaller, Ulrike Ulmen, Patrick a Acquah, Ekkehart Dietz, Teunis a Eggelte, and Ulrich Bienzle. Detection and clinical manifestation of placental malaria in southern Ghana. *Malaria journal*, 5:119, January 2006. ISSN 1475-2875. doi: 10.1186/1475-2875-5-119. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1716171&tool=pmcentrez&rendertype=abstract.

[107] Asher Mullard. Malaria: Sticking around. *Nature Reviews Microbiology*, 5 (11):831–831, November 2007. ISSN 1740-1526. doi: 10.1038/nrmicro1785. URL http://www.nature.com/doifinder/10.1038/nrmicro1785.

[108] Jo Erika T Narciso, Iris Diana C Uy, April B Cabang, Jenina Faye C Chavez, Juan Lorenzo B Pablo, Gisela P Padilla-Concepcion, and Eduardo a Padlan. Analysis of the antibody structure based on high-resolution crystallographic studies. *New biotechnology*, 28(5): 435–47, September 2011. ISSN 1876-4347. doi: 10.1016/j.nbt.2011.03.012. URL http://www.ncbi.nlm.nih.gov/pubmed/21477671.

[109] Jacques Neefjes, Marlieke L M Jongsma, Petra Paul, and Oddmund Bakke. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature reviews. Immunology*, 11(12):823–36, December 2011. ISSN 1474-1741. doi: 10.1038/nri3084. URL http://www.ncbi.nlm.nih.gov/pubmed/22076556.

[110] Morten a Nielsen, Vera V Pinto, Mafalda Resende, Madeleine Dahlbäck, Sisse B Ditlev, Thor G Theander, and Ali Salanti. Induction of adhesion-inhibitory antibodies against placental Plasmodium falciparum parasites by using single domains of VAR2CSA. *Infection and immunity*, 77(6):2482–7, June 2009. ISSN 1098-5522. doi: 10.1128/IAI.00159-09. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2687338&tool=pmcentrez&rendertype=abstract.

[111] Nancy A Obuchowski. Receiver operating characteristic curves and their use in radiology. *Radiology*, 229(1):3–8, October 2003. ISSN 0033-8419. doi: 10.1148/radiol.2291010898. URL http://www.ncbi.nlm.nih.gov/pubmed/14519861.

[112] Benjamin Ollomo, Patrick Durand, Franck Prugnolle, Emmanuel Douzery, Céline Arnathau, Dieudonné Nkoghe, Eric Leroy, and François Renaud. A new malaria agent in African hominids. *PLoS pathogens*, 5(5): e1000446, May 2009. ISSN 1553-7374. doi: 10.1371/journal.ppat.1000446. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2680981&tool=pmcentrez&rendertype=abstract.

[113] Michael Otto. Staphylococcus colonization of the skin and antimicrobial peptides. *Expert review of dermatology*, 5(2):183–195, April 2010. ISSN 1746-9872. doi: 10.1586/edm.10.6. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2867359&tool=pmcentrez&rendertype=abstract.

[114] Peter Parham. MHC class I molecules and KIRs in human history, health and survival. *Nature reviews. Immunology*, 5(3):201–14, March 2005. ISSN 1474-1733. doi: 10.1038/nri1570. URL http://www.ncbi.nlm.nih.gov/pubmed/15719024.

[115] Noa D Pasternak and Ron Dzikowski. PfEMP1: an antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite Plasmodium falciparum. *The international journal of biochemistry & cell biology*, 41(7):1463–6, July 2009. ISSN 1878-5875. doi: 10.1016/j.biocel.2008.12.012. URL http://www.ncbi.nlm.nih.gov/pubmed/19150410.

[116] Susan K Pierce. Lipid rafts and B-cell activation. *Nature reviews. Immunology*, 2(2):96–105, February 2002. ISSN 1474-1733. doi: 10.1038/nri726. URL http://www.ncbi.nlm.nih.gov/pubmed/11910900.

[117] Franck Prugnolle, Patrick Durand, Cécile Neel, Benjamin Ollomo, Francisco J Ayala, Céline Arnathau, Lucie Etienne, Eitel Mpoudi-Ngole, Dieudonné Nkoghe, Eric Leroy, Eric Delaporte, Martine Peeters, and François Renaud. African great apes are natural hosts of multiple related malaria species, including Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America*, 107(4): 1458–63, January 2010. ISSN 1091-6490. doi: 10.1073/pnas.0914440107. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2824423&tool=pmcentrez&rendertype=abstract.

[118] Anthony W Purcell, James McCluskey, and Jamie Rossjohn. More than one reason to rethink the use of peptides in vaccine design. *Nature reviews. Drug discovery*, 6(5):404–14, May 2007. ISSN 1474-1776. doi: 10.1038/nrd2224. URL http://www.ncbi.nlm.nih.gov/pubmed/17473845.

[119] Nicolas Rapin, Ilka Hoof, Ole Lund, and Morten Nielsen. MHC motif viewer. *Immunogenetics*, 60(12):759–65, December 2008. ISSN 1432-1211. doi: 10.1007/s00251-008-0330-2. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2613509&tool=pmcentrez&rendertype=abstract.

[120] Rino Rappuoli, Christian W Mandl, Steven Black, and Ennio De Gregorio. Vaccines for the twenty-first century society. *Nature reviews. Immunology*, 11(12):865–72, December 2011. ISSN 1474-1741. doi: 10.1038/nri3085. URL http://www.ncbi.nlm.nih.gov/pubmed/22051890.

[121] Luc Reininger, Miguel Garcia, Andrew Tomlins, Sylke Müller, and Christian Doerig. The Plasmodium falciparum, Nima-related kinase Pfnek-4: a marker for asexual parasites committed to sexual differentiation. *Malaria journal*, 11:250, January 2012. ISSN 1475-2875. doi: 10.1186/1475-2875-11-250. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3495404&tool=pmcentrez&rendertype=abstract.

[122] Stephen M Rich, Fabian H Leendertz, Guang Xu, Matthew LeBreton, Cyrille F Djoko, Makoah N Aminake, Eric E Takang, Joseph L D Diffo, Brian L Pike, Benjamin M Rosenthal, Pierre Formenty, Christophe Boesch, Francisco J Ayala, and Nathan D Wolfe. The origin of malignant malaria. *Proceedings of the National Academy of Sciences of the United States of America*, 106(35):14902–7, September 2009. ISSN 1091-6490. doi: 10.1073/pnas.0907740106. URL http://www.pnas.org/content/106/35/14902http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2720412&tool=pmcentrez&rendertype=abstract.

[123] Stefan Riedel. Edward Jenner and the history of smallpox and vaccination. *Proceedings (Baylor University. Medical Center)*, 18(1):21–5, January 2005. ISSN 0899-8280. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1200696&tool=pmcentrez&rendertype=abstract.

[124] Eleanor M Riley and V Ann Stewart. Immune mechanisms in malaria: new insights in vaccine development. *Nature medicine*, 19(2):168–78, March 2013. ISSN 1546-170X. doi: 10.1038/nm.3083. URL http://www.ncbi.nlm.nih.gov/pubmed/23389617.

[125] D J Roberts, A G Craig, A R Berendt, R Pinches, G Nash, K Marsh, and C I Newbold. Rapid switching to multiple antigenic and adhesive phenotypes in malaria. *Nature*, 357(6380):689–92, June 1992. ISSN 0028-0836. doi: 10.1038/357689a0. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3731710&tool=pmcentrez&rendertype=abstract.

[126] Mario H Rodriguez and Fidel De La C Hernández-Hernández. Insect-malaria parasites interactions: the salivary gland. *Insect biochemistry and molecular biology*, 34(7):615–24, July 2004. ISSN 0965-1748. doi: 10.1016/j.ibmb.2004.03.014. URL http://www.ncbi.nlm.nih.gov/pubmed/15242702.

[127] Bernard Rosner. *Fundamentals of Biostatistics*. Brooks/Cole, Boston, MA 02210, 7th edition, 2010. ISBN 9780538733496.

[128] Jeffrey Sachs and Pia Malaney. The economic and social burden of malaria. *Nature*, 415(6872):680–5, February 2002. ISSN 0028-0836. doi:

10.1038/415680a. URL
http://www.ncbi.nlm.nih.gov/pubmed/11832956.

[129] Ali Salanti, Madeleine Dahlbäck, Louise Turner, Morten a Nielsen, Lea Barfod, Pamela Magistrado, Anja T R Jensen, Thomas Lavstsen, Michael F Ofori, Kevin Marsh, Lars Hviid, and Thor G Theander. Evidence for the involvement of VAR2CSA in pregnancy-associated malaria. *The Journal of experimental medicine*, 200(9):1197–203, November 2004. ISSN 0022-1007. doi: 10.1084/jem.20041579. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2211857&tool=pmcentrez&rendertype=abstract.

[130] Ali Salanti, Mafalda Resende, Sisse B Ditlev, Vera V Pinto, Madeleine Dahlbäck, Gorm Andersen, Tom Manczak, Thor G Theander, and Morten a Nielsen. Several domains from VAR2CSA can induce Plasmodium falciparum adhesion-blocking antibodies. *Malaria journal*, 9:11, January 2010. ISSN 1475-2875. doi: 10.1186/1475-2875-9-11. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2817698&tool=pmcentrez&rendertype=abstract.

[131] Robert Sallares. *Malaria and Rome A History of Malaria in Ancient Italy*. Oxford University Press, 2002. ISBN 0199248508.

[132] Elisabeth Sappenfield, Denise J Jamieson, and Athena P Kourtis. Pregnancy and susceptibility to infectious diseases. *Infectious diseases in obstetrics and gynecology*, 2013:752852, January 2013. ISSN 1098-0997. doi: 10.1155/2013/752852. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3723080&tool=pmcentrez&rendertype=abstract.

[133] Robert W Sauerwein, Meta Roestenberg, and Vasee S Moorthy. Experimental human challenge infections can accelerate clinical malaria vaccine development. *Nature reviews. Immunology*, 11(1):57–64, January 2011. ISSN 1474-1741. doi: 10.1038/nri2902. URL http://www.ncbi.nlm.nih.gov/pubmed/21179119.

[134] B Schittek, R Hipfel, B Sauer, J Bauer, H Kalbacher, S Stevanovic, M Schirle, K Schroeder, N Blin, F Meier, G Rassner, and C Garbe. Dermcidin: a novel human antibiotic peptide secreted by sweat glands. *Nature immunology*, 2(12):1133–7, December 2001. ISSN 1529-2908.

doi: 10.1038/ni732. URL http://www.ncbi.nlm.nih.gov/pubmed/11694882.

[135] Louis Schofield and Georges E Grau. Immunological processes in malaria pathogenesis. *Nature reviews. Immunology*, 5(9):722–35, September 2005. ISSN 1474-1733. doi: 10.1038/nri1686. URL http://www.ncbi.nlm.nih.gov/pubmed/16138104.

[136] Inessa Schwab and Falk Nimmerjahn. Intravenous immunoglobulin therapy: how does IgG modulate the immune system? *Nature reviews. Immunology*, 13(3):176–89, March 2013. ISSN 1474-1741. doi: 10.1038/nri3401. URL http://www.ncbi.nlm.nih.gov/pubmed/23411799.

[137] CE Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(July, October):379–423, 623–656, 1948. URL http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

[138] Ganesh N Sivalingam and Adrian J Shepherd. An analysis of B-cell epitope discontinuity. *Molecular immunology*, 51(3-4):304–9, July 2012. ISSN 1872-9142. doi: 10.1016/j.molimm.2012.03.030. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3657695&tool=pmcentrez&rendertype=abstract.

[139] Geoffrey L Smith and Grant McFadden. Smallpox: anything to declare? *Nature reviews. Immunology*, 2(7):521–7, July 2002. ISSN 1474-1733. doi: 10.1038/nri845. URL http://www.ncbi.nlm.nih.gov/pubmed/12094226.

[140] Joseph D Smith and Kirk W Deitsch. Pregnancy-associated malaria and the prospects for syndrome-specific antimalaria vaccines. *The Journal of experimental medicine*, 200(9):1093–7, November 2004. ISSN 0022-1007. doi: 10.1084/jem.20041974. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2211864&tool=pmcentrez&rendertype=abstract.

[141] Hergen Spits and Tom Cupedo. Innate lymphoid cells: emerging insights in development, lineage relationships, and function. *Annual review of immunology*, 30:647–75, January 2012. ISSN 1545-3278. doi: 10.1146/annurev-immunol-020711-075053. URL http://www.ncbi.nlm.nih.gov/pubmed/22224763.

[142] Danielle I Stanisic, Alyssa E Barry, and Michael F Good. Escaping the immune system: How the malaria parasite makes vaccine development a challenge. *Trends in parasitology*, 29(12):612–22, December 2013. ISSN 1471-5007. doi: 10.1016/j.pt.2013.10.001. URL http://www.ncbi.nlm.nih.gov/pubmed/24176554.

[143] Alexandra Minna Stern and Howard Markel. The history of vaccines and immunization: familiar patterns, new challenges. *Health affairs (Project Hope)*, 24(3):611–21, 2005. ISSN 0278-2715. doi: 10.1377/hlthaff.24.3.611. URL http://www.ncbi.nlm.nih.gov/pubmed/15886151.

[144] Mary M Stevenson and Eleanor M Riley. Innate immunity to malaria. *Nature reviews. Immunology*, 4(3):169–80, March 2004. ISSN 1474-1733. doi: 10.1038/nri1311. URL http://www.ncbi.nlm.nih.gov/pubmed/15039754.

[145] Stuart G Tangye and David M Tarlinton. Memory B cells: effectors of long-lived immune responses. *European journal of immunology*, 39(8): 2065–75, August 2009. ISSN 1521-4141. doi: 10.1002/eji.200939531. URL http://www.ncbi.nlm.nih.gov/pubmed/19637202.

[146] Mahamadou a Thera and Christopher V Plowe. Vaccines for malaria: how close are we? *Annual review of medicine*, 63:345–57, January 2012. ISSN 1545-326X. doi: 10.1146/annurev-med-022411-192402. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3338248&tool=pmcentrez&rendertype=abstract.

[147] Stuart E Turvey and David H Broide. Innate immunity. *The Journal of allergy and clinical immunology*, 125(2 Suppl 2):S24–32, March 2010. ISSN 1097-6825. doi: 10.1016/j.jaci.2009.07.016. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2832725&tool=pmcentrez&rendertype=abstract.

[148] Yeung L Tutterrow, Marion Avril, Kavita Singh, Carole a Long, Robert J Leke, Grace Sama, Ali Salanti, Joseph D Smith, Rose G F Leke, and Diane W Taylor. High levels of antibodies to multiple domains and strains of VAR2CSA correlate with the absence of placental malaria in Cameroonian women living in an area of high Plasmodium falciparum transmission. *Infection and immunity*, 80(4):1479–90, April 2012. ISSN

1098-5522. doi: 10.1128/IAI.00071-12. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3318421&tool=pmcentrez&rendertype=abstract.

[149] Jeffrey B Ulmer and Margaret a Liu. Ethical issues for vaccines and immunization. *Nature reviews. Immunology*, 2(4):291–6, May 2002. ISSN 1474-1733. doi: 10.1038/nri780. URL http://www.ncbi.nlm.nih.gov/pubmed/12002000.

[150] M H Van Regenmortel. Antigenicity and immunogenicity of synthetic peptides. *Biologicals : journal of the International Association of Biological Standardization*, 29(3-4):209–13, 2001. ISSN 1045-1056. doi: 10.1006/biol.2001.0308. URL http://www.ncbi.nlm.nih.gov/pubmed/11851317.

[151] Marc H V Van Regenmortel. What is a B-cell epitope? *Methods in molecular biology (Clifton, N.J.)*, 524:3–20, January 2009. ISSN 1064-3745. doi: 10.1007/978-1-59745-450-6_1. URL http://link.springer.com/10.1007/978-1-59745-450-6http://link.springer.com/protocol/10.1007/978-1-59745-450-6_1http://www.ncbi.nlm.nih.gov/pubmed/19377933.

[152] Swaminathan Venkatesh, Jerry L Workman, Mats Wahlgren, and Maria Teresa Bejarano. Malaria: Molecular secrets of a parasite. *Nature*, 499(7457):156–7, July 2013. ISSN 1476-4687. doi: 10.1038/nature12407. URL http://www.ncbi.nlm.nih.gov/pubmed/23823720.

[153] Jennifer a Walker, Jillian L Barlow, and Andrew N J McKenzie. Innate lymphoid cells–how did we miss them? *Nature reviews. Immunology*, 13 (2):75–87, February 2013. ISSN 1474-1741. doi: 10.1038/nri3349. URL http://www.ncbi.nlm.nih.gov/pubmed/23292121.

[154] Elizabeth Ann Winzeler. Malaria research in the post-genomic era. *Nature*, 455(7214):751–6, October 2008. ISSN 1476-4687. doi: 10.1038/nature07361. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705782&tool=pmcentrez&rendertype=abstract.

[155] World Health Organization. World Malaria Report 2013. Technical report, World Health Organization, 2013. URL http://www.who.int/iris/bitstream/10665/97008/1/9789241564694_eng.pdf.

[156] Jianying Yang and Michael Reth. Oligomeric organization of the B-cell antigen receptor on resting cells. *Nature*, 467(7314):465–9, September 2010. ISSN 1476-4687. doi: 10.1038/nature09357. URL http://www.ncbi.nlm.nih.gov/pubmed/20818374.

[157] Maria Yazdanbakhsh and David L Sacks. Why does immunity to parasites take so long to develop? *Nature reviews. Immunology*, 10(2):80–1, February 2010. ISSN 1474-1741. doi: 10.1038/nri2673. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3437742&tool=pmcentrez&rendertype=abstract.

[158] Jonathan W Yewdell, Eric Reits, and Jacques Neefjes. Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nature reviews. Immunology*, 3(12):952–61, December 2003. ISSN 1474-1733. doi: 10.1038/nri1250. URL http://www.ncbi.nlm.nih.gov/pubmed/14647477.

[159] Zhong-Wei Zhang, Jian Cheng, Fei Xu, Yang-Er Chen, Jun-Bo Du, Ming Yuan, Feng Zhu, Xiao-Chao Xu, and Shu Yuan. Red blood cell extrudes nucleus and mitochondria against oxidative stress. *IUBMB life*, 63(7):560–5, July 2011. ISSN 1521-6551. doi: 10.1002/iub.490. URL http://www.ncbi.nlm.nih.gov/pubmed/21698761.

5 Supplementary Materials

5.1 Part II - Paper I: SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments

Published online 12 June 2013 *Nucleic Acids Research, 2013, Vol. 41, No. Web Server issue* **1–5**

SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments - Supplementary Materials

TABLE OF CONTENTS

THE SIGNISITE METHOD

The initial step of *SigniSite* is to check if the submitted set of sequences is aligned. This is done by checking the length of each sequence. If not all sequences are of the same length, i.e. not aligned, a multiple sequence alignment (MSA) will be created using MAFFT with accurate options ('mafft-einsi') (1).

To perform the *SigniSite* analysis, each sequence must have an associated real number, quantifying the phenotype of the dataset. The sequence associated real number must be placed, white-space separated, at the end of each FASTA header in the MSA. If this is not the case, *SigniSite* will assume that the submitted sequences are pre-sorted with respect to some desired phenotype (The web-server *will* alert the user if pre-sorting is assumed). The values can be sorted either ascending or descending (default). If using ascending sorting, the lowest value(s), is considered the strongest, e.g. binding affinity. If using descending sorting, the highest value(s), is considered the strongest, e.g. fluorescent label intensity. *SigniSite* utilises a non-parametric approach for the analysis, in that *SigniSite* will perform the analysis based on the ranks of the sequence associated real numbers, rather than the values of these. In the following the details of the *SigniSite* method will be elaborated.

Rank matrix generation

The first step of the *SigniSite* method is to sort the submitted MSA with respect to the sequence-associated real values.

Let N, $(N \geq 2)$ be the number of sequences in the MSA and $n_{p,a}$, $\left(1 \leq n_{p,a} < N\right)$ be the number of a specific amino acid type a observed at a specific position p in the MSA. Henceforth subscript 'p,a' will denote amino acid residue type a at position p in the MSA. Initially the sequences in the MSA are sorted descending on their associated real numbers. We can now assign a rank value to each sequence, so that the first sequence gets a rank of one, the second a rank of two, etc. In case two or more sequences share the same annotated real number value, the sequences are assigned the mean of the ranks they occupy. Each type of amino acid residue a observed at position p is subsequently assigned the rank of the sequence they appear in. Given these rank values, we can for each position p in the MSA and for each type of amino acid residue a observed at p calculate an observed mean rank value as:

$$\overline{x}_{p,a}^{obs} = \frac{1}{n_{p,a}} \sum_{i=1}^{N} rank_{p,b_i} \cdot \delta(b_i, a) \quad \textbf{(1)}$$

where the sum is over all sequences in the MSA and b_i is the amino acid at position p in the $i'th$ sequence in the MSA, such that $\delta(b_i,a) = 1$ if $b_i = a$ and $b_i,a = 0$ if $b_i \neq a$. The result of this is a mxn_{aa} rank matrix, R, where the number of rows, m, is the number of positions in the MSA and the number of columns, $n_{aa} = 20$, corresponds to the 20 proteogenic amino acids, sorted according to 'A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V'. Each position in R, $r_{i,j}$, $1 \leq i \leq m$, $1 \leq j \leq n_{aa}$ hold the mean of the assigned ranks for amino acid residue a at position p.

Please note that since each sequence as prerequisite has one sequence associated value, each sequence, and subsequently each amino acid at each position, has an assigned rank. The number of sequences, sequence associated values and ranks are thus all equal to N.

The statistical framework of *SigniSite*, *z-score* calculation

The statistical framework of *SigniSite*, is such that the null-hypothesis for the non-parametric statistical test that is performed by *SigniSite* is: H_0: *Amino acid residue a at position p has no preference for 'strong' or 'weak' phenotypic values* and the alternative hypothesis is: H_1: *Amino acid residue a at position p has a significant preference for either 'strong' or 'weak' phenotypic values*, (two-tailed test) i.e.:

$$H_0: \quad \mu_{p,a}^{exp} = \overline{x}_{p,a}^{obs} \quad H_1: \quad \mu_{p,a}^{exp} \neq \overline{x}_{p,a}^{obs} \quad \textbf{(2)}$$

Where $\mu_{p,a}^{exp}$ is the expected mean of the ranks and $\overline{x}_{p,a}^{obs}$ is the observed mean rank. Under the null-hypothesis, we can then compute a standard score $z_{p,a}$ quantifying the probability of observing $\overline{x}_{p,a}^{obs}$:

$$z_{p,a} = \frac{\mu_{p,a}^{exp} - \overline{x}_{p,a}^{obs}}{\sigma_{p,a}^{exp}} \quad \textbf{(3)}$$

where $\sigma_{p,a}^{exp}$, $\sigma_{p,a}^{exp} > 0$ is the standard deviation of the mean expected rank given the composition of amino acid residues

at the position in question. Based on the magnitude of the $z_{p,a}$, we can then compare with the level of significance and subsequently either reject or accept the null-hypothesis.

$\mu_{p,a}^{exp}$ and $\sigma_{p,a}^{exp}$ could be computed by reshuffling (permuting) the residues present at a given position p a large number of times. For larger data sets, this approach however becomes computationally unfeasible. The values are therefore more efficiently estimated using an analytical approximation. The statistical evaluation performed by *SigniSite* is similar to the Wilcoxon rank statistic (2), and we can therefore analytically derive approximations for $\mu_{p,a}^{exp}$ and $\sigma_{p,a}^{exp}$. For $\mu_{p,a}^{exp}$ this yields:

$$\mu_{p,a}^{exp} = \frac{N+1}{2} \quad (4)$$

recall that N is the number of sequences in the MSA. For $\sigma_{p,a}^{exp}$ we get:

$$\sigma_{p,a}^{exp} = \sqrt{\frac{(N-n_{p,a})(N+1)\cdot t_c}{12\cdot n_{p,a}}} \quad (5)$$

recall that $n_{p,a}$ is the number of occurrences of residue a at position p. If a given position in the MSA is fully conserved, i.e. $n_{p,a}=N \Rightarrow \sigma_{p,a}^{exp}=0$. In this case, the amino acid a at position p is assigned $z_{p,a}=0$. t_c, $0<t_c\leq 1$, is the tie-correction factor. $t_c=1\Rightarrow$ no tied values, $t_c=0\Rightarrow$ only tied values (not allowed, since $t_c=0\Rightarrow\sigma_{p,a}^{exp}=0$, for which $z_{p,a}$ is not defined). The tie-correction factor adjusts for tied observations and is computed by defining a tie-vector, **T**, where each element $t_1,t_2,...,t_h$ (h being the number of unique sequence-associated values) is the count of occurrences of a given value v_i (2). The tie-correction factor t_c is defined as:

$$t_c = 1 - \frac{1}{N^3-N}\cdot\sum_{i=1}^{h}(t_i^3-t_i) \quad (6)$$

Given a random set of amino acid sequences and associated numerical values, the distribution of assigned *z-scores* at a position Z_p can be approximated by the normal distribution $Z_p \sim \mathcal{N}(\mu=0,\sigma^2=1)$, thus allowing straightforward assignment of *p-values* to each of the observations.

The final result is a *z-score* matrix, with the same dimensions as the rank matrix.

Correction for Multiple Comparisons

SigniSite will perform one test per residue per position in the MSA. Clearly, this raises a multiple testing scenario, as the more hypotheses we test, the higher the chance that we obtain at least one false positive result. Based on the assigned *p-value*, the user can address the multiple testing problem by two different methods: Bonferroni's single-step and Holm's step-down procedure (3). Bonferroni correction is more conservative than Holm correction. A detailed elaboration of these procedures is beyond the scope of this study and the reader is referred to Dudoit *et al.,* 2002 for details on these procedures.

Example of calculations

The following is a simple fictive example for illustrating how to perform an evaluation. In an alignment, at p_{13}, we observe 'R' in 5 of 20 sequences. After descending sorting, 'R' occupies ranks 1,2,5,6,7. We now know the ranks, that $n_{13,R}=5$ and that $N=20$, therefore:

$$\overline{x}_{13,R}^{mean} = \frac{1}{5}\sum_{i=1}^{20} rank_{13,b_i}\cdot\delta(b_i,R) = \frac{1}{5}\cdot(1+2+5+6+7) = 4.2$$

$$\sigma_{13,R}^{exp} = \sqrt{\frac{(20-5)(20+1)\cdot 1}{12\cdot 5}} = 2.3$$

$$\mu_{13,R}^{exp} = \frac{20+1}{2} = 10.5$$

$$z_{p,a} = \frac{10.5-4.2}{2.3} = 2.75$$

The final $z_{13,R}=2.75$ corresponds to an uncorrected *p-value* of $p_{13,R}=0.006$. At first this may seem highly signficant, but if e.g. a total of 50 tests are performed when analysing the alignment, the Bonferroni corrected *p-value* becomes:

$$p_{adj}^{Bonf.} = min(1, p\cdot n_{tests}) = min(1, 0.006\cdot 50) = 0.3$$

Corresponding to $z_{adj}=1.04$.

BENCHMARK DATASETS

Stanford University HIVdb Genotype-Phenotype Datasets

Human immunodeficiency virus type 1 (HIV-1) Protease Genotype-Phenotype Datasets (GPDs) (Version 5.0, March, 2012) were downloaded from the Stanford University HIV Drug Resistance Database (HIVdb) (4, 5), available at [http://HIVdb.stanford.edu/cgi-bin/GenoPhenoDS.cgi]. The GPDs consist of sequenced variants of the HIV-1 protease, where the fold-change in resistance of each variant (compared to wild-type) against 8 different Protease Inhibitors (PIs), namely *APV, ATV, IDV, LPV, NFV, RTV, SQV, TPV* (See section "Overview of HIV-1 PIs" for details) has been measured using 3 different assays ('Antivirogram' (Virco^{TM}), 'PhenoSense' (ViroLogic^{TM}) and 'All Others').

Compilation of HIVdb multiple sequence alignments

Multiple sequence alignments (MSAs) were compiled from the GPDs. Each MSA contains the sequences of a set of HIV-1 protease variants, each with measured fold change in resistance (compared to wild-type) towards the *same* PI, measured using the *same* assay. The sequences were filtered such that any sequences containing: *'no sequence' (.), 'insertions' (#), 'deletions' (∼), 'stop codon' (*) or 'unknown amino acid residue' (X)* (according to HIVdb nomenclature) were excluded. We noted that these exclusions reduced the size of the data, we however deemed high-confidence data to be the more important parameter. At positions where HIVdb states that two or more residues were observed (i.e. a mixture) a random of the observed residues was selected, such that only *one* sequence per sequence id was constructed (see section: "The impact of chosen seed for random number generation" for details). It should be noted that the choice only to construct one sequence per sequence id, was deliberate due to combinatorics. The HIV-1 protease consists of 99 residues, if just 5 positions held a mixture of 3 residues the outcome of all combinations would be 243 sequences, thus greatly skewing the MSA towards this/these variants. The assayed PIs listed in the GPDs were cross-referenced with those in the information in table of Resistance Mutation Scores (RMS), and the intersection of PIs available for analysis and comparison was: *ATV, IDV, LPV, NFV, SQV and TPV* each of which was measured using the 'Antivirogram' (Virco^{TM}), 'PhenoSense' (ViroLogic^{TM}) and 'All Others' types of assays. A total of 12,714 sequence distributed on 18 multiple sequence alignments (MSAs) were compiled this way (6 PIs times 3 assays) (see Table 1).

Observed discrepancies in the HIVdb

We noted that in the GPD 'Antivirogram', the sequence id '159250' contained a lowercase 'i' in the 'P32' column and in the GPD 'PhenoSense' the sequence id '45124' contained a ',' in the 'IsolateName' column.

RMS: HIVdb Resistance Mutation Scores

The RMS table was downloaded from the HIVdb PI Resistance Notes (4, 5), available at [http://HIVdb.stanford.edu/DR/cgi-bin/rules_scores_HIVdb.cgi?class=PI]. The RMS table contains information about positions with known resistance mutations and their individual impact on the resistance towards 8 different PIs compared to wild-type. The scores range from -10 to 60, where a positive score indicates that this particular mutation away from wild-type increases the resistance towards a given PI. A negative score in turn indicates that there is an increase in susceptibility towards the PI (i.e. decreased resistance). At each position in the table of RMS harbouring a resistance mutation, the consensus residue was assigned a RMS of 0.

Table 1. Summary of the number of sequences in each of the 18 multiple sequence alignments (MSAs) used in the benchmark of *SigniSite*. One MSA was compiled for each set of HIV-1 protease sequences with a fold change in resistance (compared to wild-type) against the *same* Protease Inhibitor (PI), measured using the *same* assay.

Assay PI	PhenoSense ViroLogic^{TM}	Antivirogram Virco^{TM}	All Others
ATV	812	670	37
IDV	1,322	1,072	235
LPV	1,097	962	15
NFV	1,374	1,104	125
SQV	1,339	1,104	242
TPV	559	632	13
Total	6,503	5,544	667

A total of 12,714 sequences were constructed. The length of each of the HIV-1 protease variants is 99 amino acid residues.

IAS: International Antiviral Society-USA - Update of the Drug Resistance Mutations in HIV-1: March 2013

From the International Antiviral Society-USAs (IAS) Update of the Drug Resistance Mutations in HIV-1: March 2013 (6), available at [https://www.iasusa.org/sites/default/files/tam/21-1-6.pdf] protease mutations known to impact PI resistance were retrieved. The annotations are given in the table: "Mutations in the protease gene associated with resistance to protease inhibitors". We assigned scores such that the residue at a given position, annotated as a resistance mutation compared to wild-type was assigned an IAS score of '1'. As with the RMS, the consensus residue at annotated resistance positions was assigned an IAS score of 0. It should be noted that the IAS scores, unlike the fold-values of the RMS, are binary.

RMS and IAS tables versions used in the benchmarking

RMS$_{bin}$: Creating the binary RMS table. The RMS downloaded from the HIVdb contained real number scores. In order to transform this table to a binary table (RMS$_{bin}$) containing actual positives (APs) and actual negatives (ANs), table scores were assigned as follows: $AP: S_{rms} > 0 \Rightarrow S_{rms_{bin}} = 1$ and $AN: S_{rms} \leq 0 \Rightarrow S_{rms_{bin}} = 0$.

(RMS+IAS)$_{mut}$: Enriching the binary RMS table with the IAS table. In order to create the enriched consensus table of RMS$_{bin}$ and IAS. Scores were assigned as follows: $AP: S_{rms_{bin}} > 0$ or $S_{ias} > 0 \Rightarrow S_{(rms+ias)_{mut}} = 1$. Otherwise $AN: S_{(rms+ias)_{mut}} = 0$.

(RMS+IAS)$_{pos}$: Positional targets for SigniSite and SPEER comparison. Positional targets were created by looking at each position annotated in the (RMS+IAS)$_{mut}$ and then assigning $AP: S_{(rms+ias)_{pos}} = 1$ if at least one score $S_{(rms+ias)_{mut}} = 1$ was found. Otherwise $AN: S_{(rms+ias)_{pos}} = 0$.

The SPEER program and SDP benchmark data

The SPEER program (Specificity Prediction using amino acids' properties, Entropy and Evolution Rate) (7, 8) was downloaded along with MSAs and corresponding experimentally annotated specificity determining sites SDS) from the SPEER repository available at: [ftp://ftp.ncbi.nih.gov/pub/SPEER/]. We downloaded the latest curated version of the data as described by Chakrabarti and Panchenko (9) (See (9) for detailed description of the

alignments). A total of 20 MSAs were downloaded, 13 of which contained only subgroup '1' and '2' assignment.

BENCHMARK STRATEGY

The strategy for benchmarking *SigniSite* followed three steps: *i*. Threshold dependent performance evaluation, *ii*. Threshold independent performance evaluation and *iii*. Comparison with existing methods.

i. Threshold dependent benchmarking

SCC$\big(z_{p,a} \sim$RMS$\big)$: The initial step in the benchmarking of *SigniSite* was to use Spearman's rank correlation (SCC) to quantify the correlation between the *SigniSite z-scores* (i.e. the $z_{p,a}$'s) obtained from the analysis of the 18 HIV-1 MSAs and the known real-number quantified fold-resistance-increase scores per mutation, as given in the RMS. If the *SigniSite z-scores* correlated with the RMS, this was an indication that *SigniSite* was able to identify resistance (phenotype) impacting mutations (genotype), i.e. perform Genotype-Phenotype correlation. The SCC was calculated at three different thresholds for including a given $z_{p,a}$: *i*. $|z_{p,a}| \geq 0 \sim$ *p-values* ≤ 1, *ii*. $|z_{p,a}| \geq 1.96 \sim$ *p-values* ≤ 0.05 and lastly *iii*. $|z_{p,a}| \geq 1.96 \sim$ *p-values* ≤ 0.05 after correction for multiple comparisons using the Bonferroni Single-Step approach, i.e. $p_{adj} = min[1, p \cdot n_{tests}]$.

[MCC, SENS, SPEC]$\big(z_{p,a} \sim$(RMS+IAS)$_{mut}\big)$: Next we calculated the standard performance measures of Sensitivity (SENS), Specificity (SPEC) and Matthew's Correlation Coefficient (MCC) letting the (RMS+IAS)$_{mut}$ define APs and ANs. These measures were calculated at three different threshold for assigning significance (Predicted Positives) to $z_{p,a}$: *i*. $|z_{p,a}| \geq 0 \sim$ *p-values* ≤ 1, *ii*. $|z_{p,a}| \geq 1.96 \sim$ *p-values* ≤ 0.05 and lastly *iii*. $|z_{p,a}| \geq 1.96 \sim$ *p-values* ≤ 0.05 after correction for multiple comparisons using the Bonferroni Single-Step approach, i.e. $p_{adj} = min[1, p \cdot n_{tests}]$.

ii. Threshold independent benchmarking

AUC$\big(z_{p,a} \sim$(RMS+IAS)$_{mut}\big)$: Having evaluated the performance of *SigniSite* using the conventional threshold dependent approaches of SCC, MCC, SENS and SPEC, we turned to the threshold independent measure of area under the receiever operator characteristics (ROC) curve (AUC). Using the *SigniSite z-scores* as prediction values and subsequently calculating the AUC$\big(z_{p,a} \sim$(RMS+IAS)$_{mut}\big)$.

iii. Comparing *SigniSite* to existing methods

Having evaluated *SigniSite* extensively on the HIVdb data sets, we turned to comparing *SigniSite* with a state-of-the-art SDP method. In a 2009 SDP benchmark review (9) the SPEER method (7, 8) was identified as the best performing method. We therefore chose to benchmark *SigniSite* against SPEER on both the HIVdb data sets and SPEER's native benchmark dataset (SDP), available at [ftp://ftp.ncbi.nih.gov/pub/SPEER].

Positional versus residue level evaluations. As SPEER makes only positional predictions, the challenge was to setup the benchmark in a fair manor, not letting any of the two methods have an advantage. As *SigniSite* makes a prediction per present residue per position, we had to transform this to a positional score. This transformation was done by assigning the maximum of the absolute *z-scores* computed per position, i.e. $z_p = max\big(|z_{p,A}|, |z_{p,R}|, ..., |z_{p,Y}|, |z_{p,V}|\big)$. The (RMS+IAS)$_{pos}$ was used for defining APs and ANs. It should be noted that *SigniSite* assigns significance to a *site* by evaluating if $max\big(|z_{p,A}|, |z_{p,R}|, ..., |z_{p,Y}|, |z_{p,V}|\big) \geq 1.96$ after CMT. As this is already build into the method we perceive the above transformation to be valid.

Scoring all positions versus excluding the unlikely. SigniSite makes at least one prediction for each position, regardless of composition, whereas SPEER will skip fully conserved positions and positons with a gap-frequency larger than 20%. We therefore assigned -100 as SPEER-score to these positions, as this was a lower score than any of the observed SPEER-scores.

SPEER requires a pre-classification of subgroups in the MSA. To make the HIVdb data sets compatible with SPEER, in that they contain only real-valued sequence-associated values, we chose to sort the HIVdb sequences descending on their associated values and subsequently split the sequences into subgroups '1' and '2' on the median of these values. If this yielded a spare sequence (uneven number of values), subgroup '1' was assigned to the spare sequence.

SigniSite requires a sequence associated real number. To make the SDP data sets compatible with *SigniSite* we chose only to include 13 of the 20 available MSAs as they included only subgroups '1' and '2' assignments. We subsequently simply assigned the subgroup number as 'real' value to each of the sequences.

SigniSite vs. SPEER: Final performance comparison Using the above framework, we computed the mean(AUC) and standard error of the mean for two sets of comparisons: *i*. *SigniSite* versus SPEER on the HIVdb data sets (18 MSAs) and *ii*. *SigniSite* versus SPEER on the SDP datasets (13 MSAs).

THE IMPACT OF CHOSEN SEED FOR RANDOM NUMBER GENERATION

At the download page for the Genotype-Phenotype Datasets from the HIVdb, it is stated that: "$p_1 ... p_n$: Two and more amino acid codes indicates a mixture". At such positions, we chose to select a random of the observed amino acid residues. In doing so, we observed that the seed used to initialise the pseudo random number generator for making a random choice of amino acids impacted the performance. In order *not* to artificially increase the performance by selecting the seed, which yielded the best performance, we instead decided to evaluate the extent of this impact. We constructed a seed-vector with $n = 1{,}000$ random integer elements $\gamma \in \big[-5^{10}, 5^{10}\big]$ using perl (v5.12.2) and a seed of -1. Each element $\gamma_1, \gamma_2, ..., \gamma_n$ was then used as seed when constructing the MSA, and the mean(SCC) were recorded as a function of the random generated seeds. We then identified the seed 4601882967 as yielding a $mean(SCC)$-performance equal

to the $mean$ of the different $mean(SCC)$-performances as a function of the $1{,}000$ random seeds. Table 2 summarises the findings.

As seen in Table 2C using a seed of 4601882967 *does not* produce results, which are significantly different from the 'true mean' as estimated by generating the 1,000 random seeds.

OVERVIEW OF HIV-1 PROTEASE INHIBITORS

Table 3 gives and overview of the protease inhibitors (PIs) mentioned in the *SigniSite* paper.

ABBREVIATIONS

Table 4 gives an overview of the abbreviations used in the *SigniSite* paper.

REFERENCES

1. Katoh,K., Misawa,K., Kuma,K.I. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**(14), 3059-3066.
PMCID: PMC135756, doi: 10.1093/nar/gkf436.
2. Armitage,P., Berry,G. and Matthews,J.N.S. (2002) *Statistical Methods in Medical Research*. Blackwell Publishing Company, Malden, MA. USA.
3. Dudoit,S., Yang,Y.H., Callow,M.J. and Speed,T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sin.*, **12**, 111-139.
LINK, PDF.
4. Rhee,S.Y., Gonzales,M.J., Kantor,R., Betts,B.J., Ravela,J., and Shafer,R.W. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.*, **30**(1), 298-303.
PMCID: PMC165547, doi: 10.1093/nar/gkg100.
5. Shafer,R.W. (2006) Rationale and Uses of a Public HIV Drug-Resistance Database. *J Infect Dis.*, **194**, S51-S58.
PMCID: PMC2614864, doi: 10.1086/505356.
6. Johnson,V.A., Calvez,V., Gnthard,H.F., Paredes,R., Pillay,D., Shafer,R., Wensing,A.M. and Richman,D.D. (2013) Update of the Drug Resistance Mutations in HIV-1: March 2013. *Top Antivir Med.*, Feb-Mar;**21**(1), 6-14.
PMID: 23596273, PDF.
7. Chakrabarti,S., Bryant,S.H. and Panchenko,A.R. (2007) Functional specificity lies within the properties and evolutionary changes of amino acids. *J Mol Biol.*, **373**, 801-810.
PMCID: PMC2605514, doi: 10.1016/j.jmb.2007.08.036.
8. Chakraborty,A., Mandloi,S., Lanczycki,C.J., Panchenko,A.R. and Chakrabarti,S. (2012) SPEER-SERVER: a web server for prediction of protein specificity determining sites. *Nucleic Acids Res.*, **40**(Web Server issue), W242-W248.
PMCID: PMC3394334, 10.1093/nar/gks559.
9. Chakrabarti,S. and Panchenko,A.R. (2009) Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC Bioinformatics*, **373**, 801-810.
PMCID: PMC2716344, doi: 10.1016/j.jmb.2007.08.036.

Table 2. Seed impact on performance. **A**: The mean of the mean(SCC) using 1,000 random seeds. **B**: The mean(SCC) when using the benchmark seed of 4601882967. **C**: The probability of observing (in A) a value as or more extreme as the obtained mean(SCC) (in B) given a seed of 4601882967. The *p-value* is calculated assuming normality in A. Quantile-quantile analysis of A gave no evidence against normal distribution (Follows the central limit theorem when sampling means).

	A 1,000 random seeds	B seed = 4601882967	C *p-value*
$\|z\| > t$	mean(mean(SCC))	mean(SCC)	B in A
$t = 0$	0.451 ± 0.00190	0.451 ± 0.0630	0.950
$t = 1.96$	0.506 ± 0.00205	0.506 ± 0.0685	0.980
$t =$ Bonf.	0.542 ± 0.00352	0.542 ± 0.0788	0.958

Rounded to 3 decimals. $|z| > t$ implies only computing the SCC for *z-scores*, with an absolute value larger than t.

Table 3. Overview of HIV-1 protease inhibitors (PIs).

Abbr.	Brand Name	Generic Name	App. Date
APV	Agenerase	Amprenavir	15-Apr-99
ATV	Reyataz	Atazanavir	20-Jun-03
DRV	Prezista	Darunavir	23-Jun-06
FPV	Lexiva	Fosamprenavir	20-Oct-03
IDV	Crixivan	Indinavir	13-Mar-96
LPV	Kaletra	Lopinavir	15-Sep-00
NFV	Viracept	Nelfinavir	14-Mar-97
RTV	Norvir	Ritonavir	1-Mar-96
SQV	Invirase	Saquinavir	6-Dec-95
TPV	Aptivus	Tipranavir	22-Jun-05

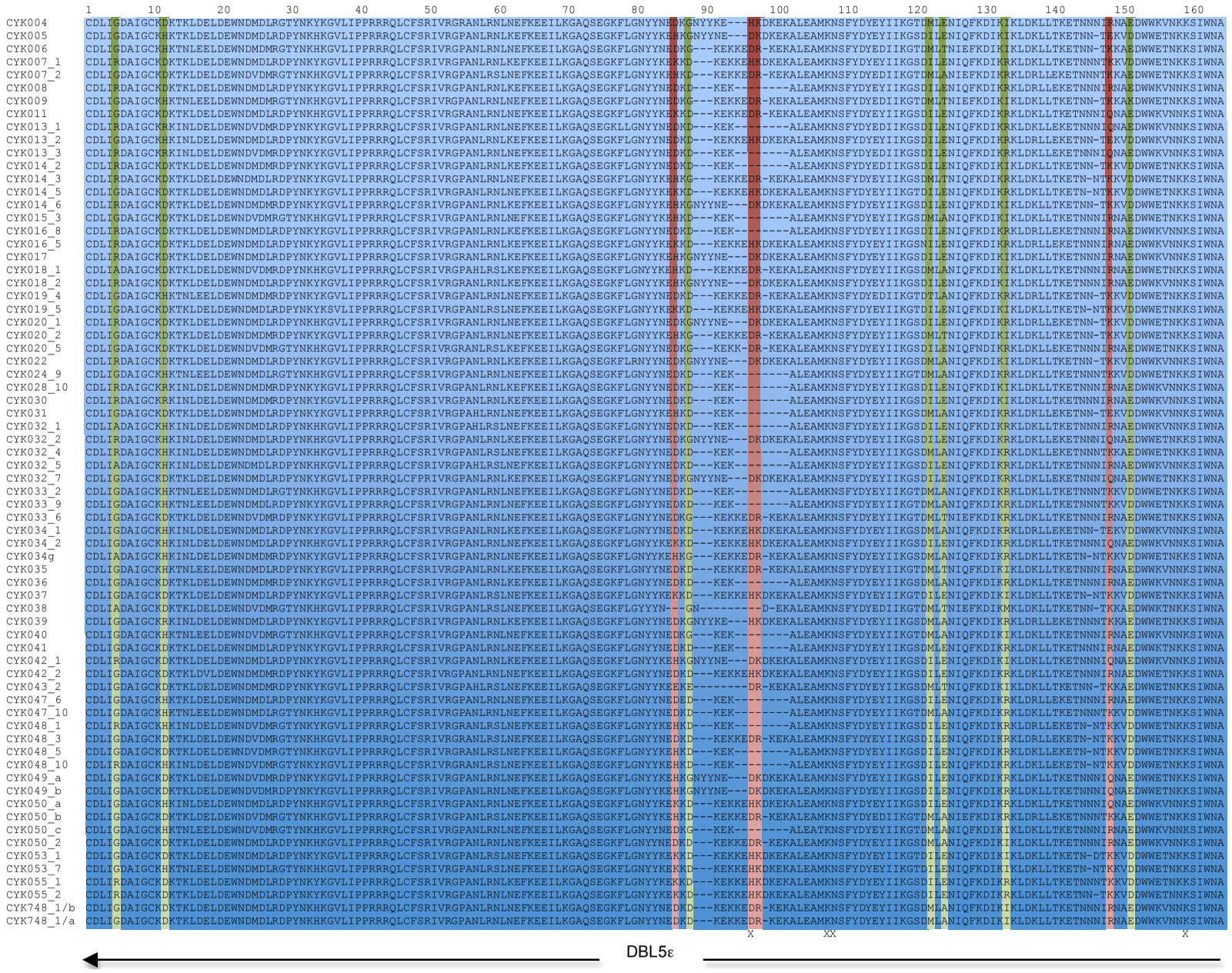
In the PI Resistance notes in the HIVDB, FPV/r, IDV/r, SQV/r, LPV/r, ATV/r, TPV/r and DRV/r are listed. the */r means in combination with RTV. Adapted from U.S. Food and Drug Administration [http://www.fda.gov/ForConsumers/byAudience/ForPatientAdvocates/HIVandAIDSActivities/ucm118915.htm] and HIVDB [http://hivdb.stanford.edu/DR/geno_clinical_review/PI.html].

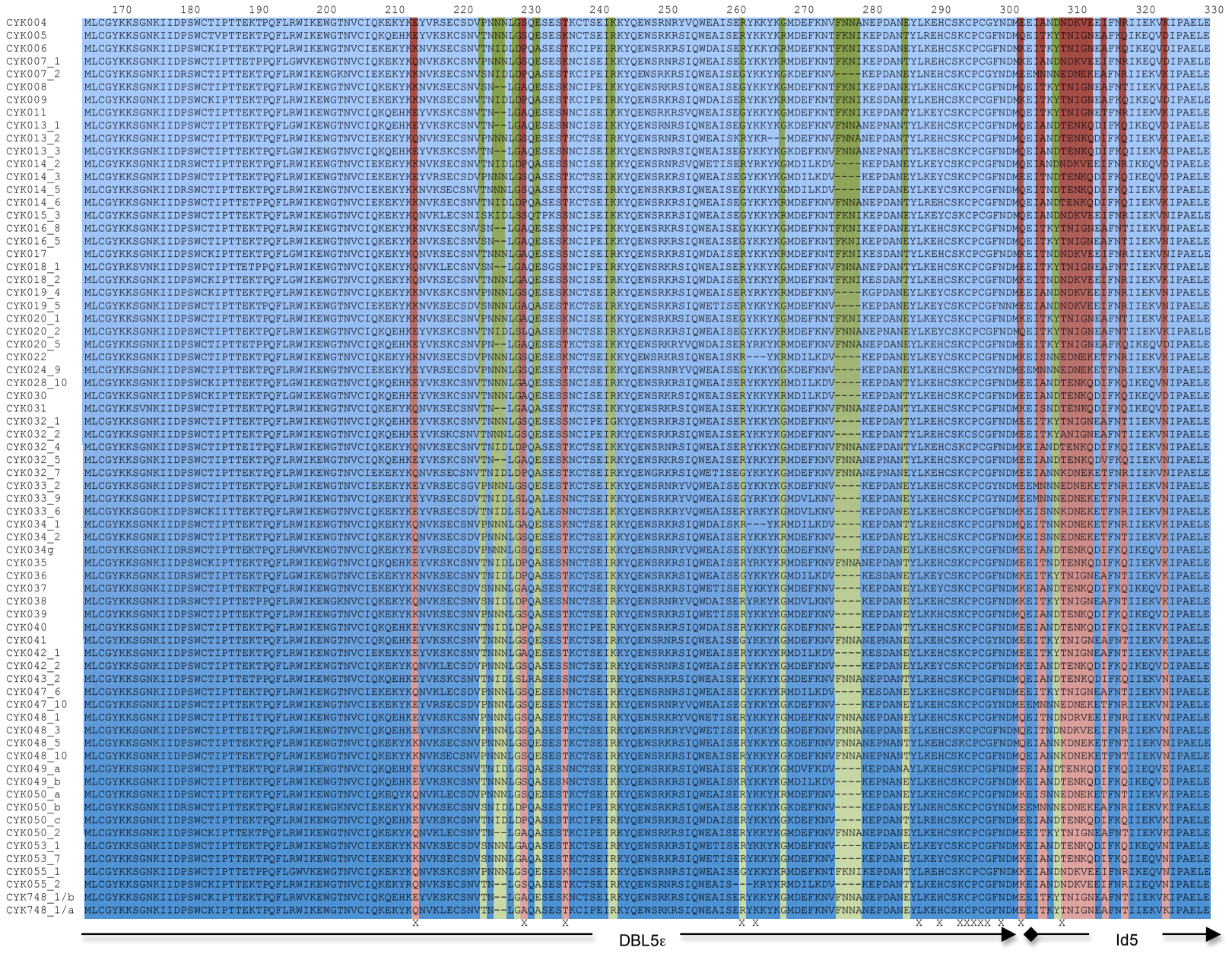
Table 4. Abbriviations used in this paper.

Abbr.	Meaning
AA	Amino Acid
AN	Actual Negative
AP	Actual Positive
APV	Amprenavir
ATV	Atazanavir
AUC	Area Under the ROC Curve
CMT	Correction for Multiple Testing
DRV	Darunavir
FPV	Fosamprenavir
GPD	Genotype-Phenotype Dataset
HIV	Human Immunodeficiency Virus
HIVdb	HIV Drug Resistance Database - Stanford University
IAS	International Antiviral Society United States of America
IDV	Indinavir
LPV	Lopinavir
MCC	Matthew's Correlation Coefficient
MSA	Multiple Sequence Alignment
NA	(data) Not Available
NFV	Nelfinavir
PI	Protease Inhibitor
ROC	Reciever Operator Characteristics
RMS	Resistance Mutation Scores
RTV	Ritonavir
SENS	Sensitivity
SCC	Spearman's rank correlation
SPEC	Specificity
SDS	Specificity Determining Site
SDP	Specificity Determining Prediction
SE	Standard Error
SPEER	Specificity Prediction using AA properties, Entropy and Evolution Rate
SQV	Saquinavir
THR	Threshold
TPV	Tipranavir
WT	Wild-type

5.2 Part III - Paper II: Insight into Antigenic Diversity of VAR2CSA-DBL5ε Domain from Multiple Plasmodium falciparum Placental Isolates

Figure of multiple alignment of parasite isolates VAR2CSA DBL5ε sequences. cDNA from 40 placental parasites isolates (39 placental isolates from Senegal and one from Tanzania) were amplified, cloned, and sequenced. Sequence ids are given at the far left. The Tanzanian isolate was isolate 748 (sequences 748_1/2a and 748_1/2b) corresponding to the DBL5? domain amplified in this isolate. The remaining sequences correspond to those obtained in isolates from Senegal. The remaining CYK are Senegalese isolates. The CYK suffix corresponds to the placenta id from which the isolate was extracted. The DBL5ε and ID5 highly conserved (blue, Shannon entropy $0 \leq H \leq 1$), conserved (green, $1 < H < 1.5$), and relatively variable (red, $1.5 \leq H \leq 2$) blocks, are indicated. The 15% most variable positions were selected and marked with 'x'.

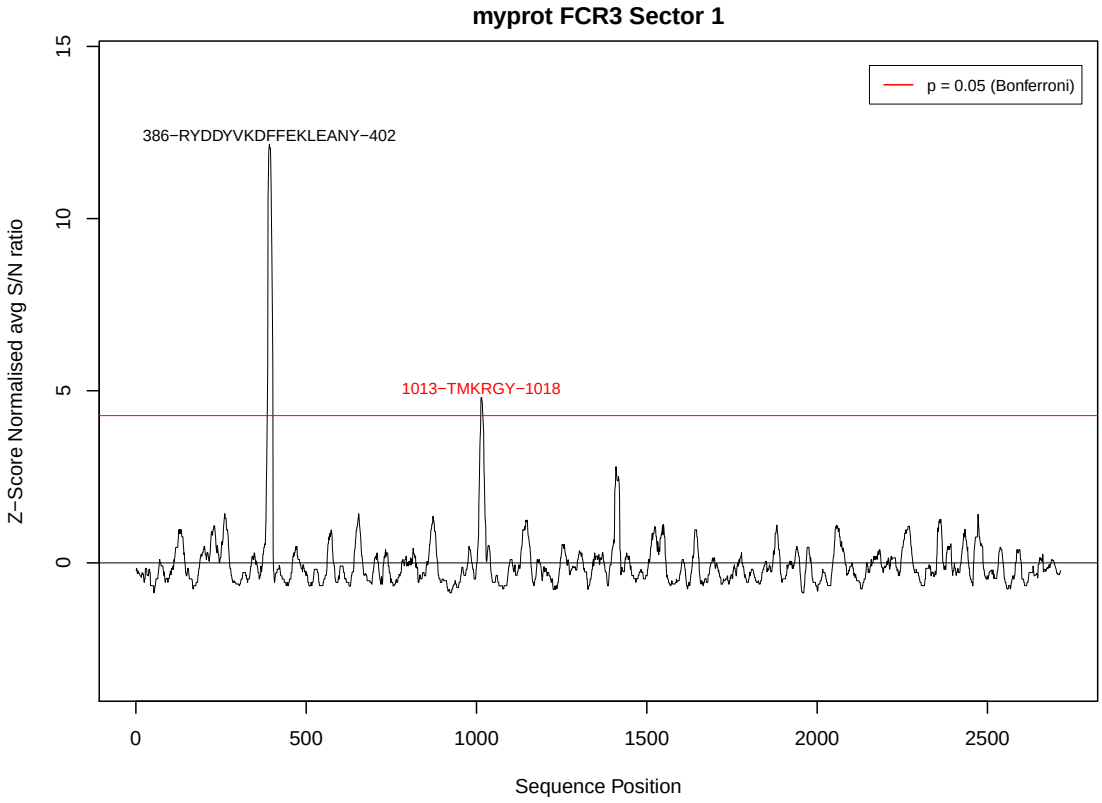

1 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
CYK004
CYK005
CYK006
CYK007_1
CYK007_2
CYK008
CYK009
CYK011
CYK013_1
CYK013_2
CYK013_3
CYK014_2
CYK014_3
CYK014_5
CYK014_6
CYK015_3
CYK016_8
CYK016_5
CYK017
CYK018_1
CYK018_2
CYK019_4
CYK019_5
CYK020_1
CYK020_2
CYK020_5
CYK022
CYK024_9
CYK028_10
CYK030
CYK031
CYK032_1
CYK032_2
CYK032_4
CYK032_5
CYK032_7
CYK033_2
CYK033_9
CYK033_6
CYK034_1
CYK034_2
CYK034g
CYK035
CYK036
CYK037
CYK038
CYK039
CYK040
CYK041
CYK042_1
CYK042_2
CYK043_2
CYK047_6
CYK047_10
CYK048_1
CYK048_3
CYK048_5
CYK048_10
CYK049_a
CYK049_b
CYK050_a
CYK050_b
CYK050_c
CYK050_2
CYK053_1
CYK053_7
CYK055_1
CYK055_2
CYK748_1/b
CYK748_1/a
X XX X
DBL5ε

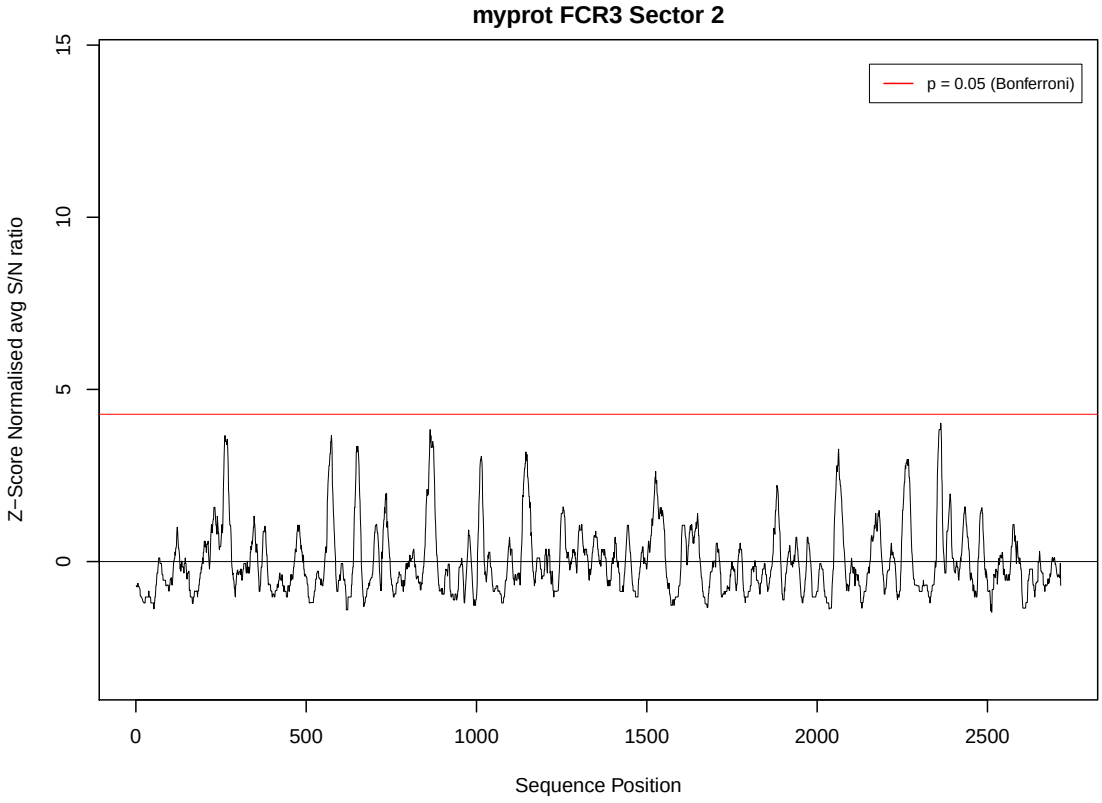

170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330
CYK004
CYK005
CYK006
CYK007_1
CYK007_2
CYK008
CYK009
CYK011
CYK013_1
CYK013_2
CYK013_3
CYK014_2
CYK014_3
CYK014_5
CYK014_6
CYK015_3
CYK016_8
CYK016_5
CYK017
CYK018_1
CYK018_2
CYK019_4
CYK019_5
CYK020_1
CYK020_2
CYK020_5
CYK022
CYK024_9
CYK028_10
CYK030
CYK031
CYK032_1
CYK032_2
CYK032_4
CYK032_5
CYK032_7
CYK033_2
CYK033_9
CYK033_6
CYK034_1
CYK034_2
CYK034g
CYK035
CYK036
CYK037
CYK038
CYK039
CYK040
CYK041
CYK042_1
CYK042_2
CYK043_2
CYK047_6
CYK047_10
CYK048_1
CYK048_3
CYK048_5
CYK048_10
CYK049_a
CYK049_b
CYK050_a
CYK050_b
CYK050_c
CYK050_2
CYK053_1
CYK053_7
CYK055_1
CYK055_2
CYK748_1/b
CYK748_1/a
DBL5ε
Id5

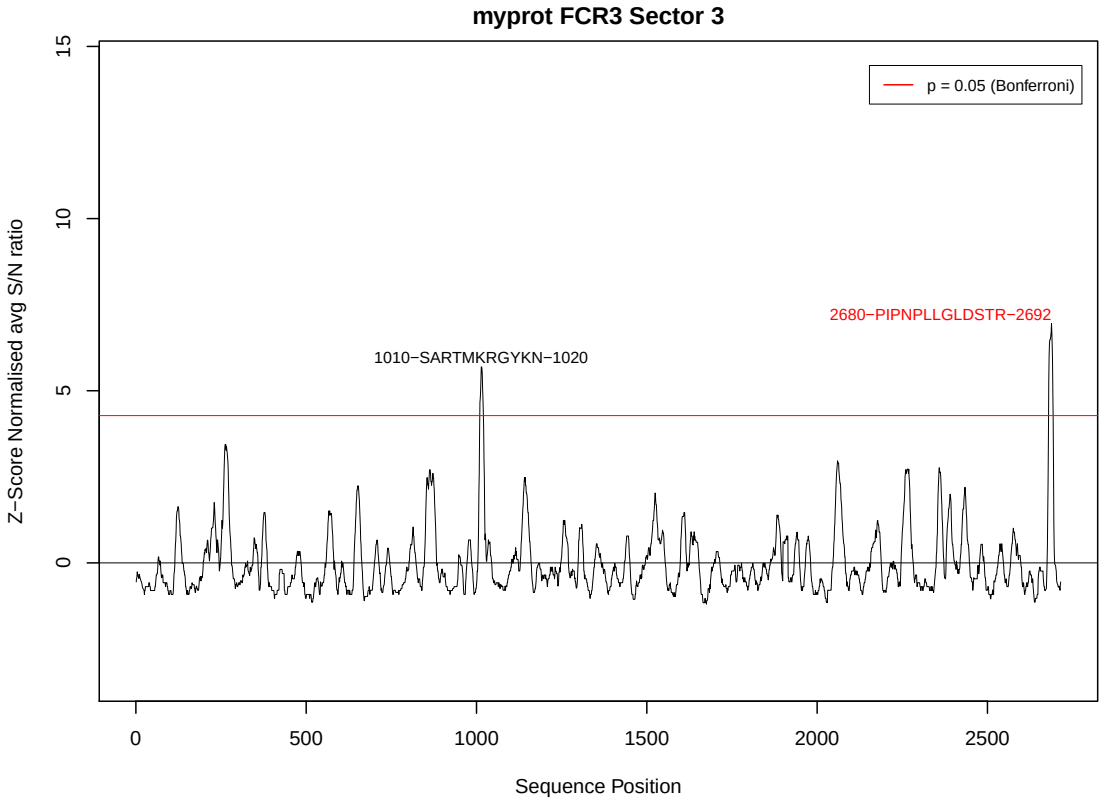
5.3 Part IV - Development and Application of Bioinformatics tool for Signal Detection in High Throughput, High Density Peptide Microarray

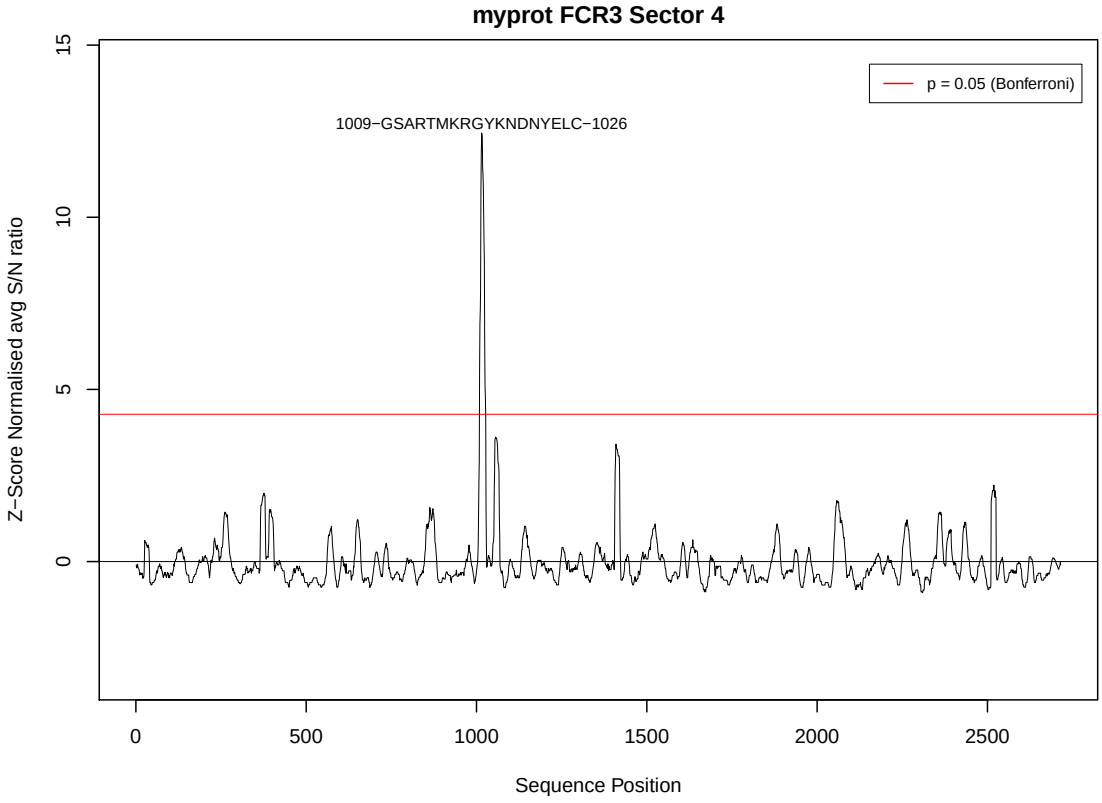
5.3.1 Complete Output from the PepChipper-1.0 Server

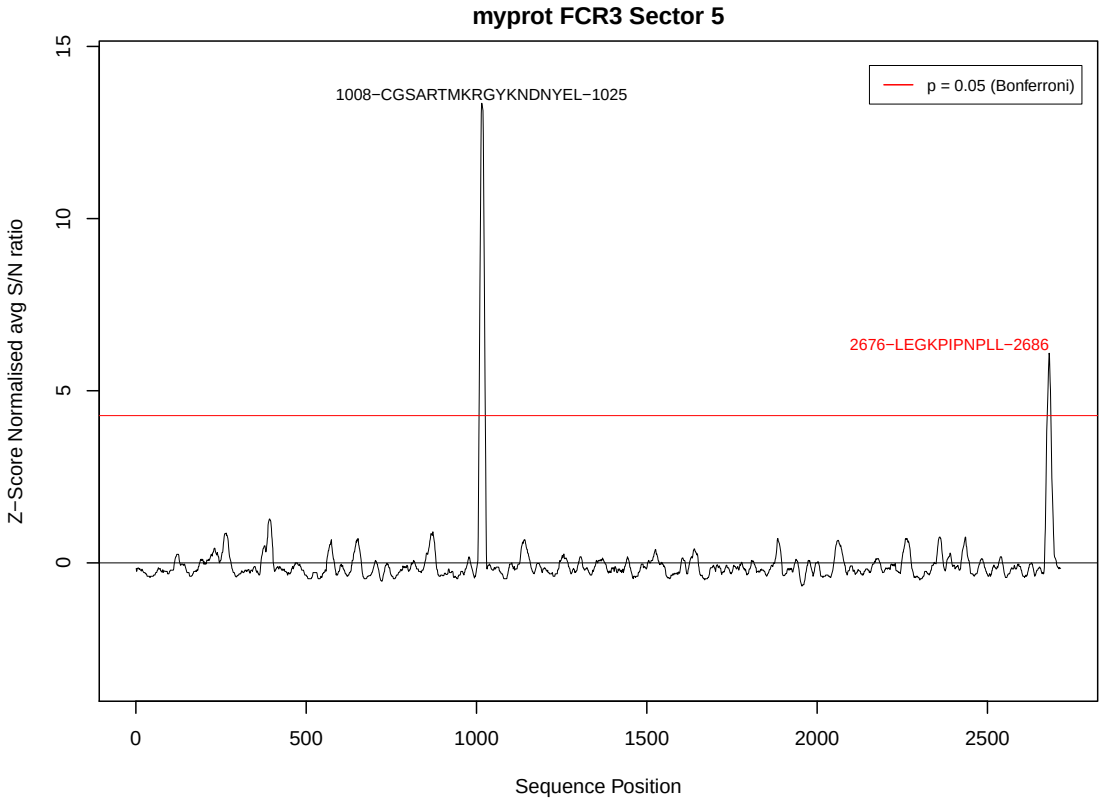
Plots of VAR2CSA FCR3 epitopes identified using the direct signal mapping approach

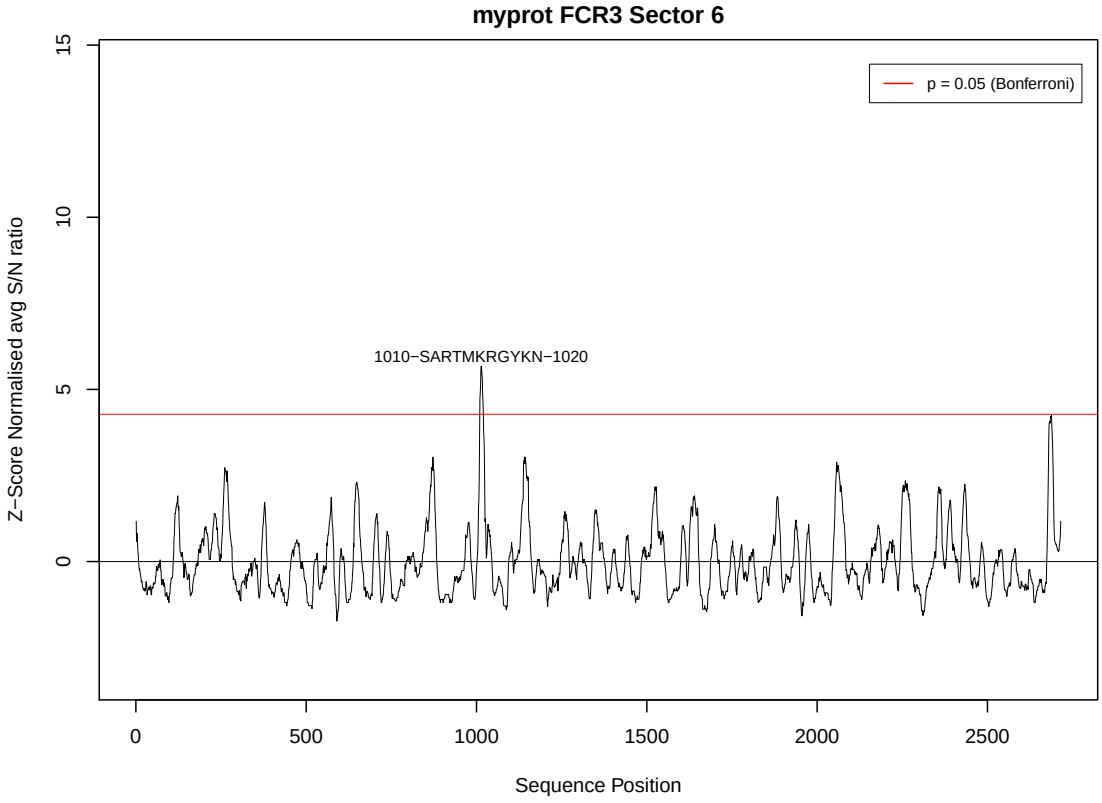

myprot FCR3 Sector 1
p = 0.05 (Bonferroni)
386−RYDDYVKDFFEKLEANY−402
1013−TMKRGY−1018
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

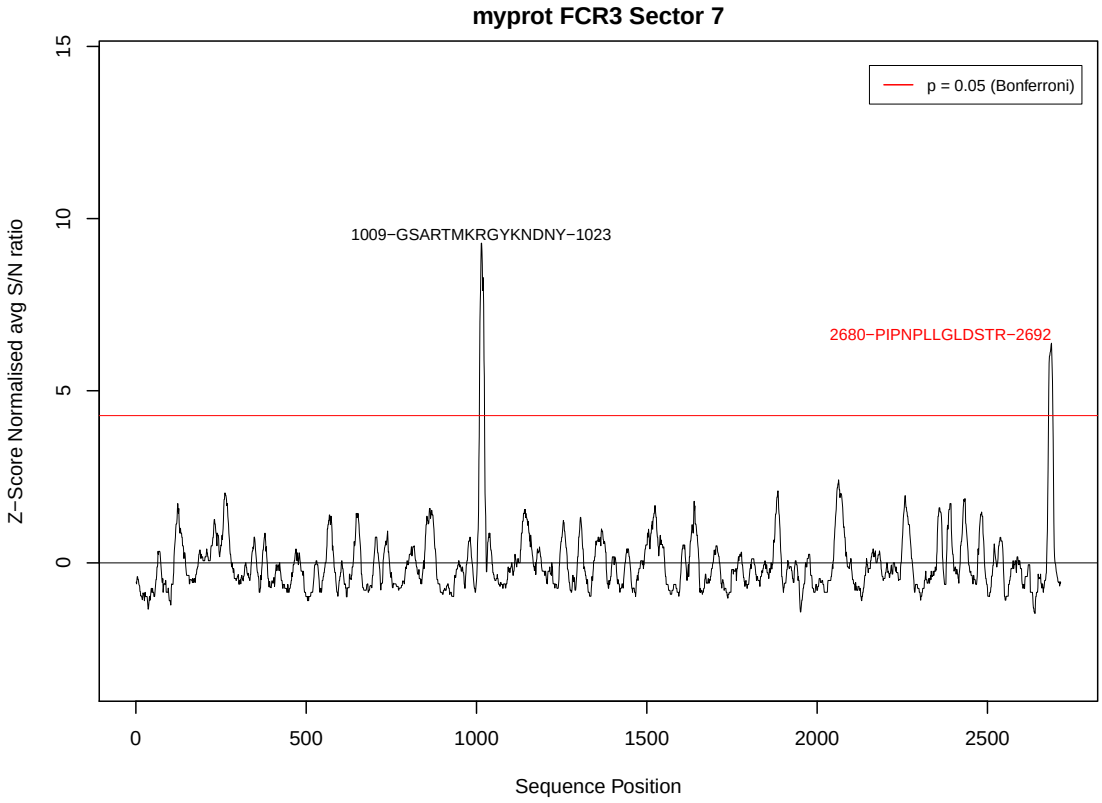

myprot FCR3 Sector 2
p = 0.05 (Bonferroni)
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

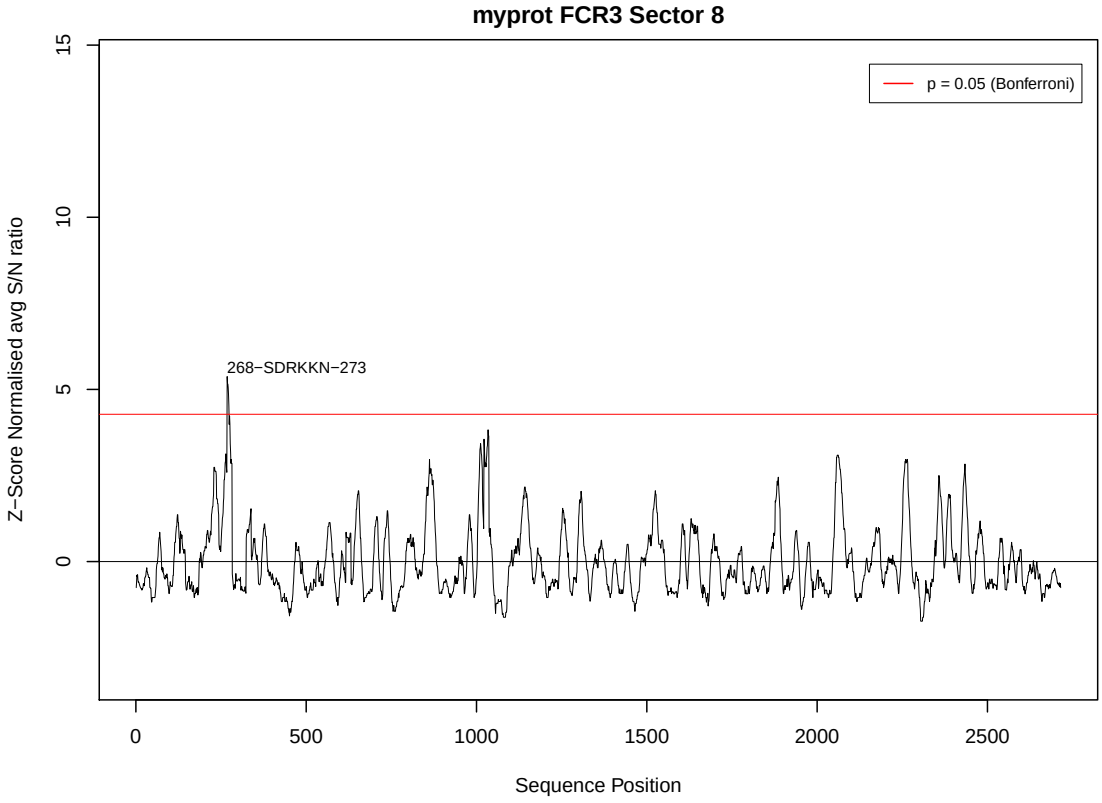

myprot FCR3 Sector 3
p = 0.05 (Bonferroni)
2680−PIPNPLLGLDSTR−2692
1010−SARTMKRGYKN−1020
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

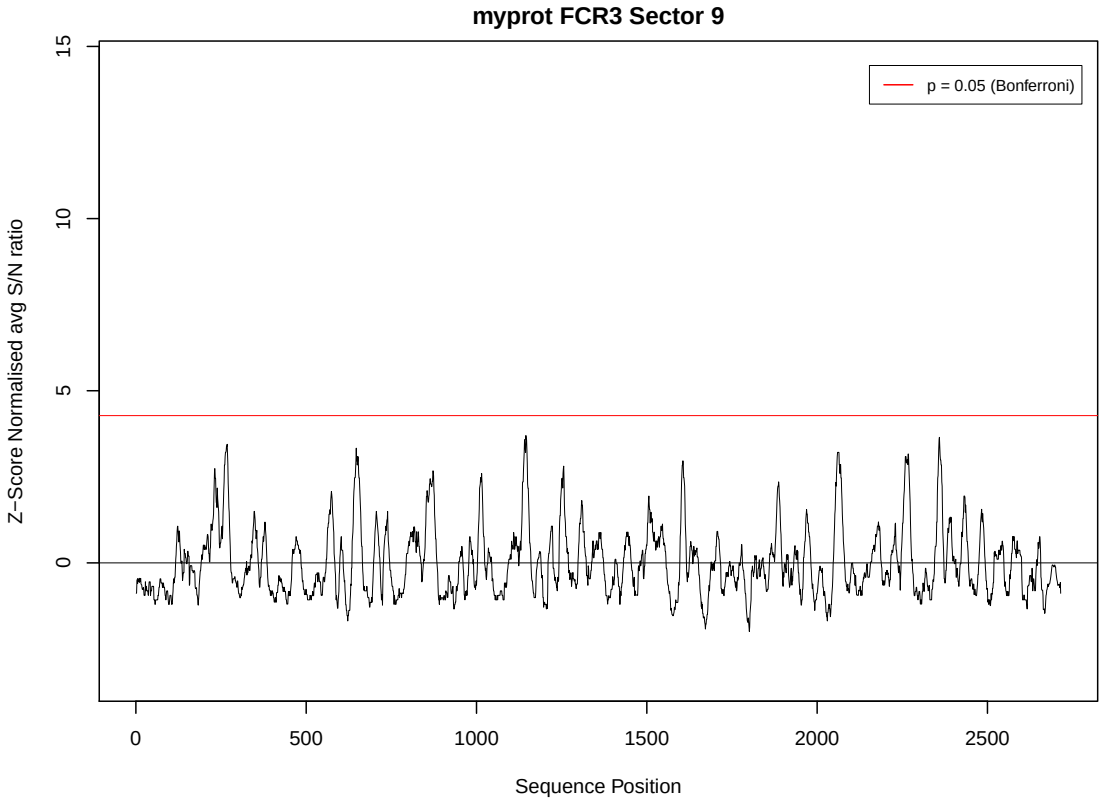

myprot FCR3 Sector 4
p = 0.05 (Bonferroni)
1009−GSARTMKRGYKNDNYELC−1026
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

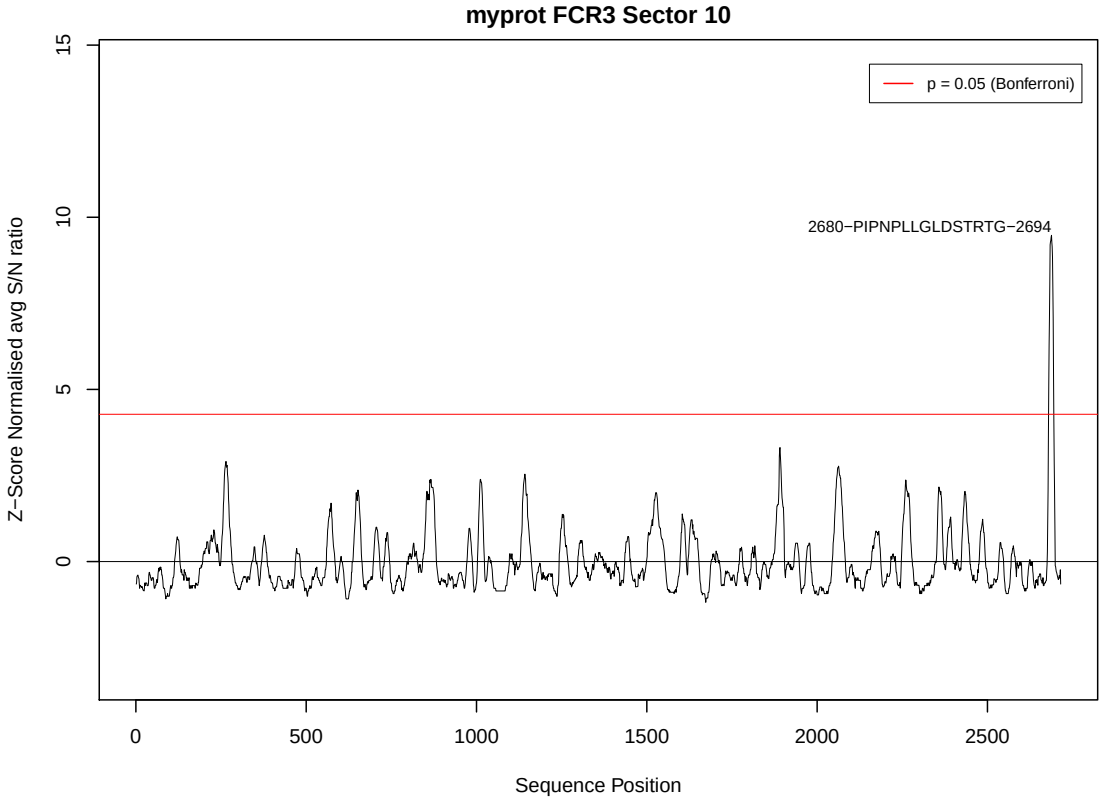

myprot FCR3 Sector 5
1008−CGSARTMKRGYKNDNYEL−1025
2676−LEGKPIPNPLL−2686
p = 0.05 (Bonferroni)
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

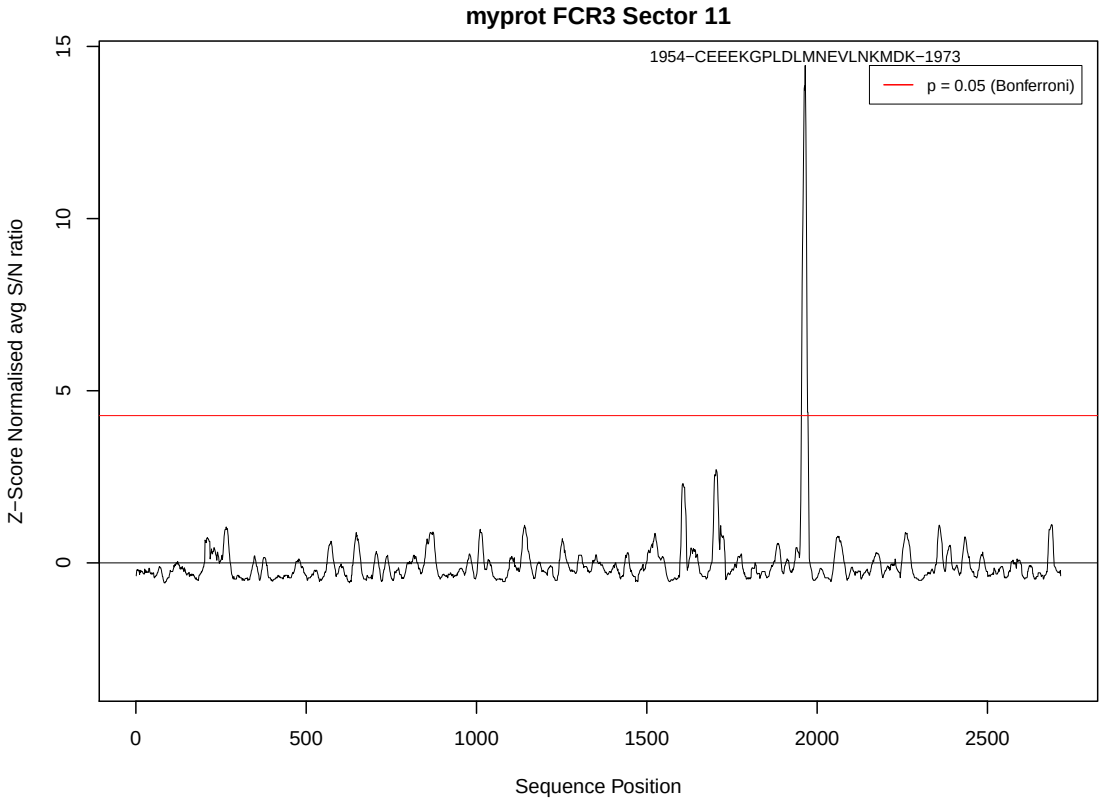

myprot FCR3 Sector 6
1010−SARTMKRGYKN−1020
p = 0.05 (Bonferroni)
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

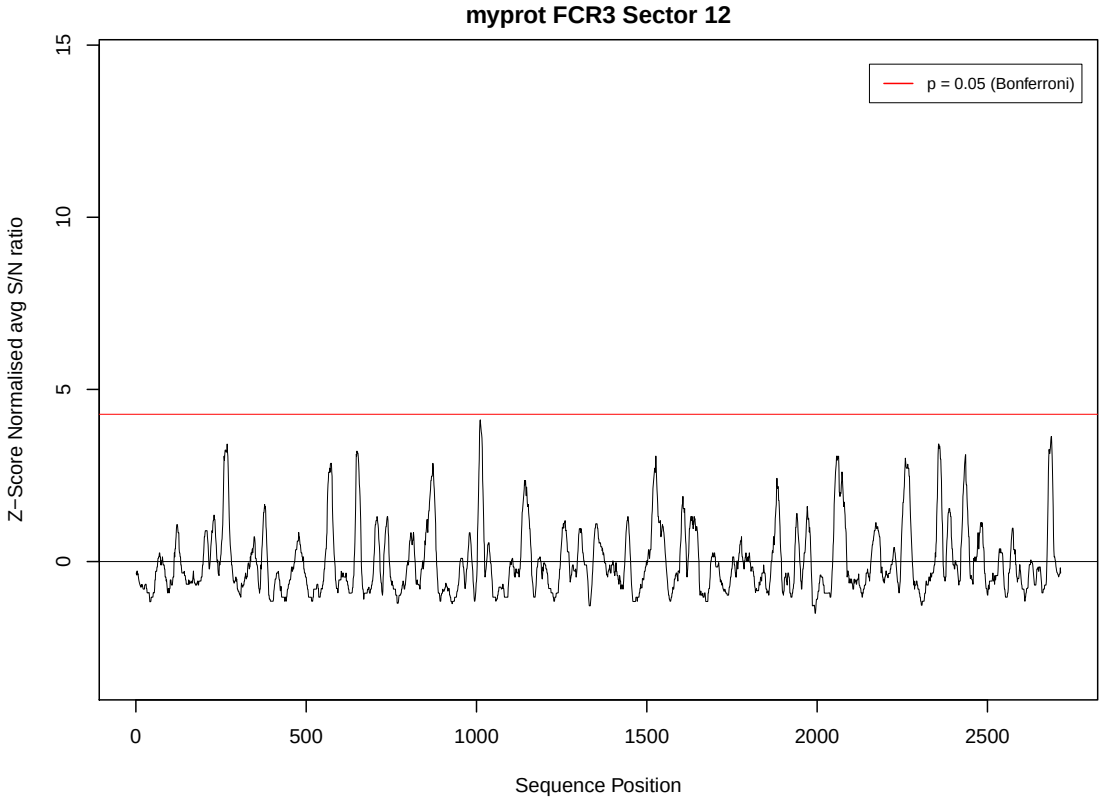

myprot FCR3 Sector 7
p = 0.05 (Bonferroni)
1009−GSARTMKRGYKNDNY−1023
2680−PIPNPLLGLDSTR−2692
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

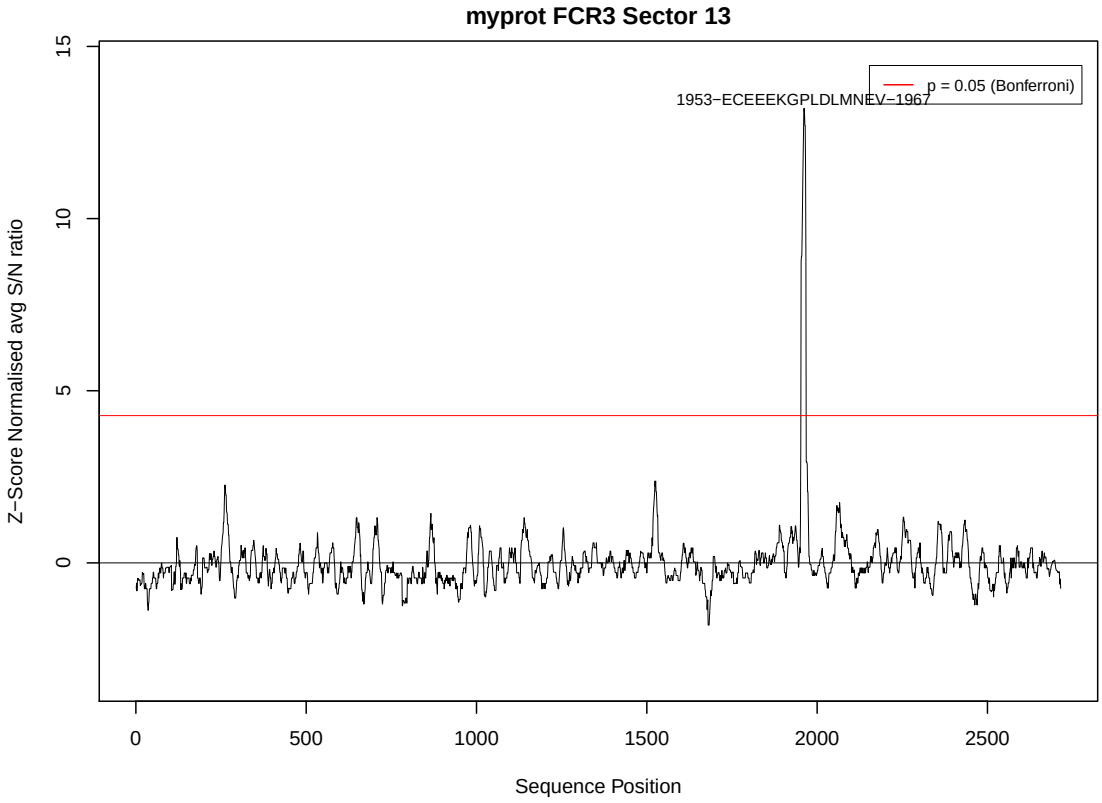

myprot FCR3 Sector 8
p = 0.05 (Bonferroni)
268−SDRKKN−273
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

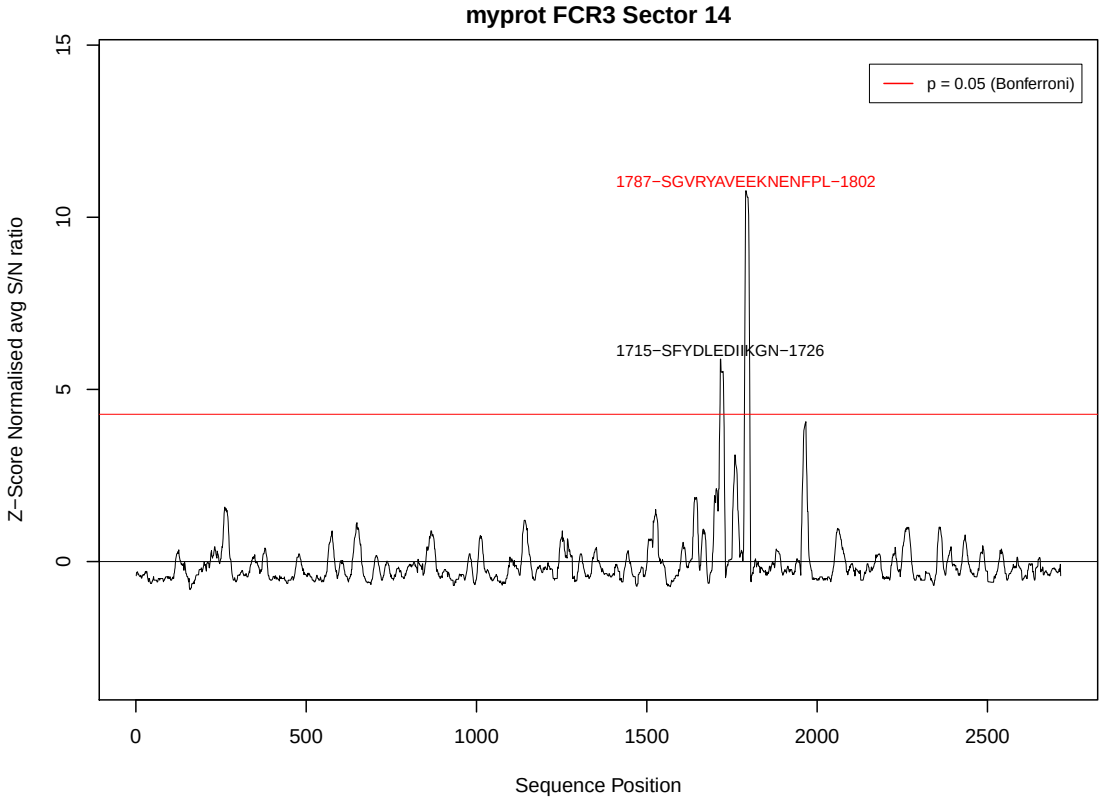

myprot FCR3 Sector 9
p = 0.05 (Bonferroni)
Z−Score Normalised avg S/N ratio
15
10
5
0
0
500
1000
1500
2000
2500
Sequence Position

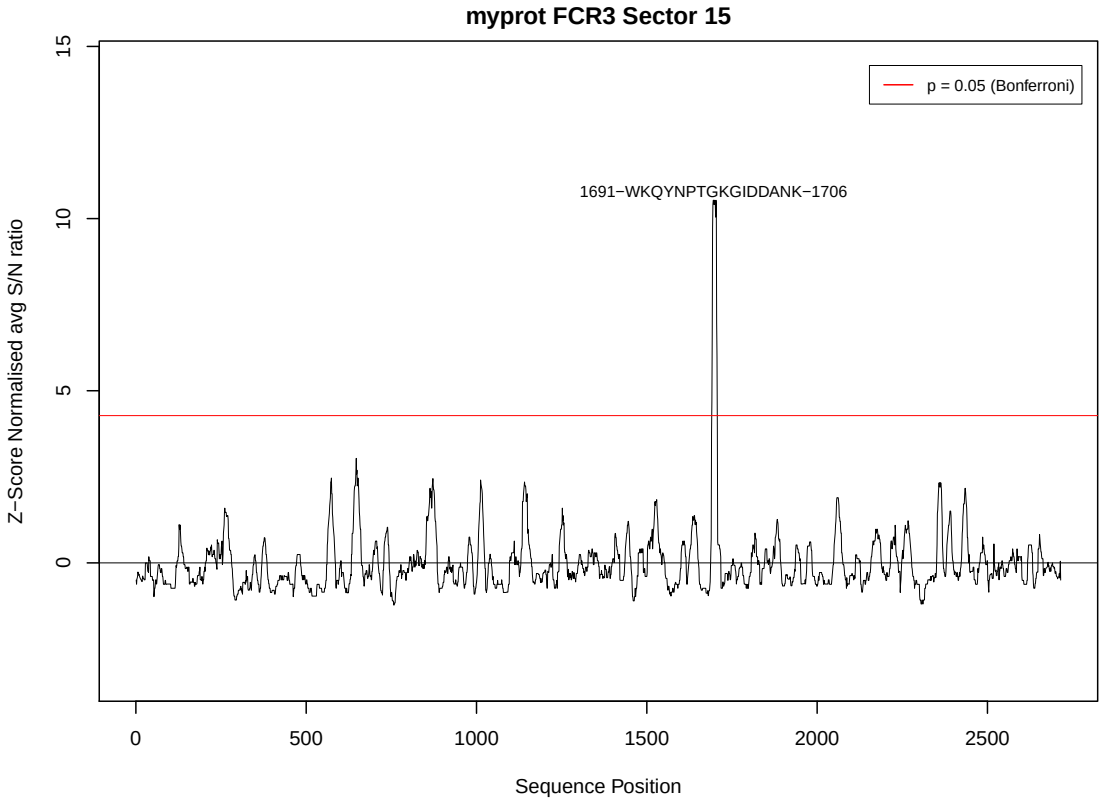

myprot FCR3 Sector 10
p = 0.05 (Bonferroni)
2680−PIPNPLLGLDSTRTG−2694
Z−Score Normalised avg S/N ratio
15
10
5
0
0
500
1000
1500
2000
2500
Sequence Position

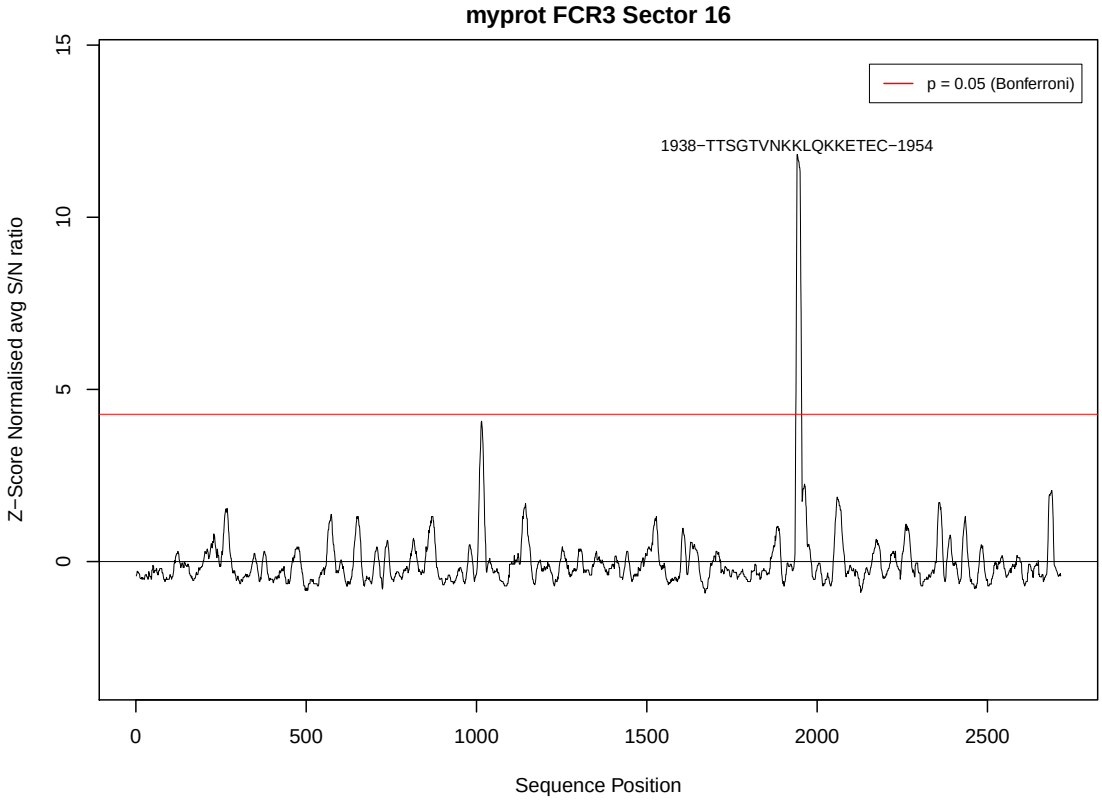

myprot FCR3 Sector 11
1954−CEEEKGPLDLMNEVLNKMDK−1973
p = 0.05 (Bonferroni)
Z−Score Normalised avg S/N ratio
0
5
10
15
0
500
1000
1500
2000
2500
Sequence Position

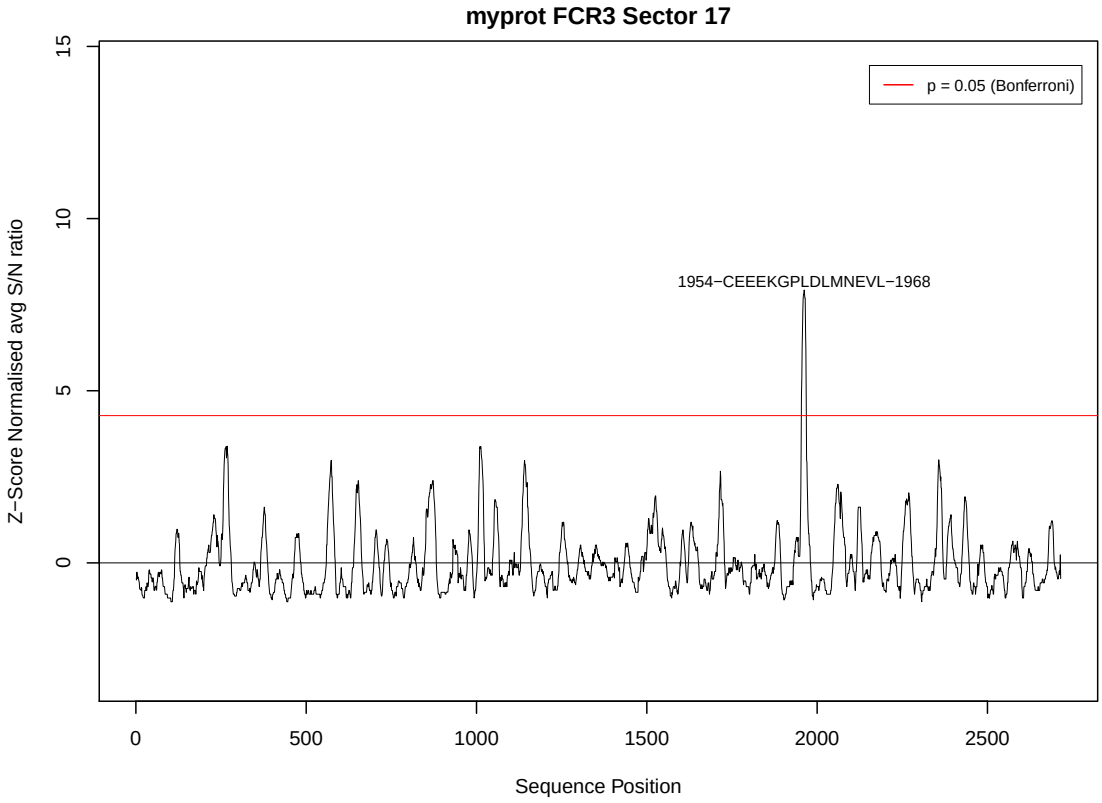

myprot FCR3 Sector 12
p = 0.05 (Bonferroni)
Z−Score Normalised avg S/N ratio
0
5
10
15
0
500
1000
1500
2000
2500
Sequence Position

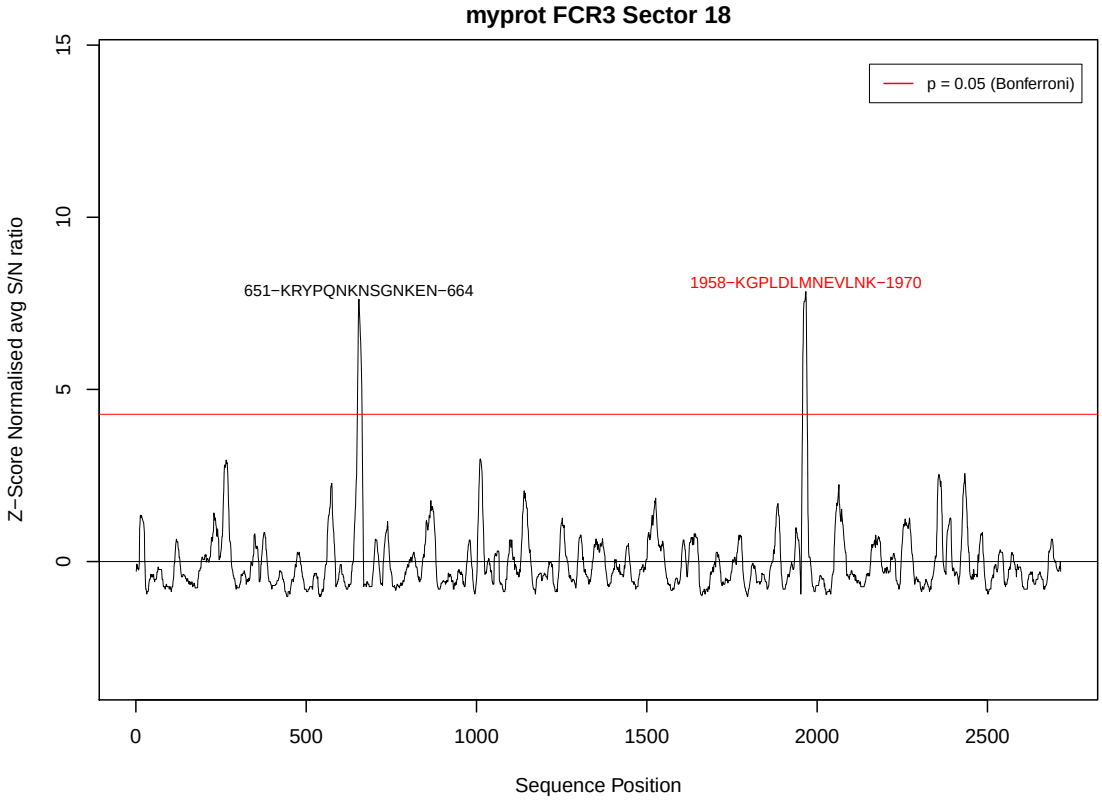

myprot FCR3 Sector 13
p = 0.05 (Bonferroni)
1953−ECEEEKGPLDLMNEV−1967
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

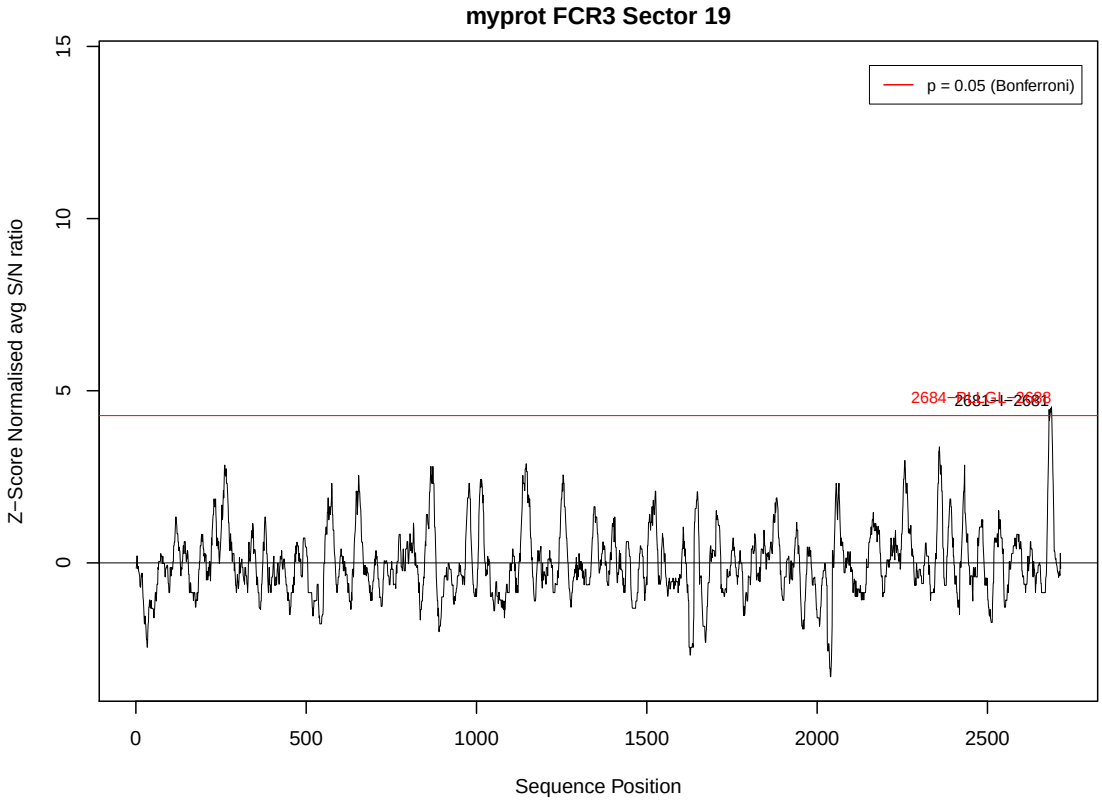

myprot FCR3 Sector 14
p = 0.05 (Bonferroni)
1787−SGVRYAVEEKNENFPL−1802
1715−SFYDLEDIIKGN−1726
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

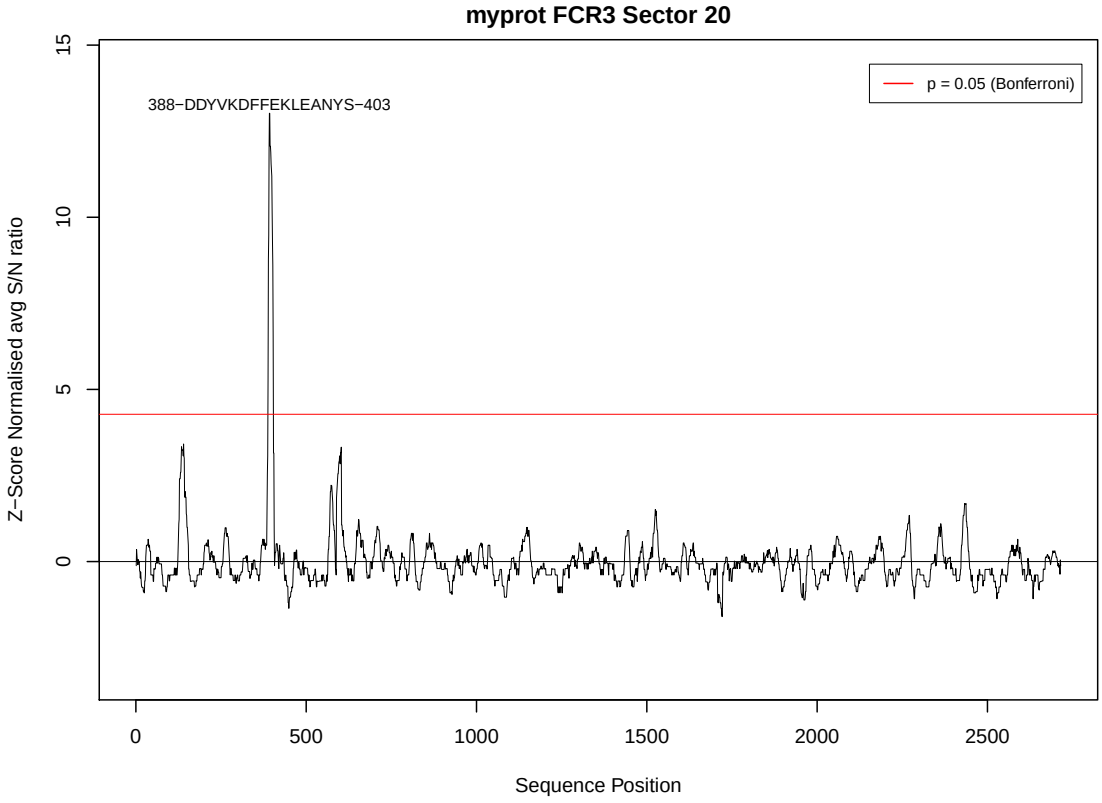

myprot FCR3 Sector 15
p = 0.05 (Bonferroni)
1691−WKQYNPTGKGIDDANK−1706
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

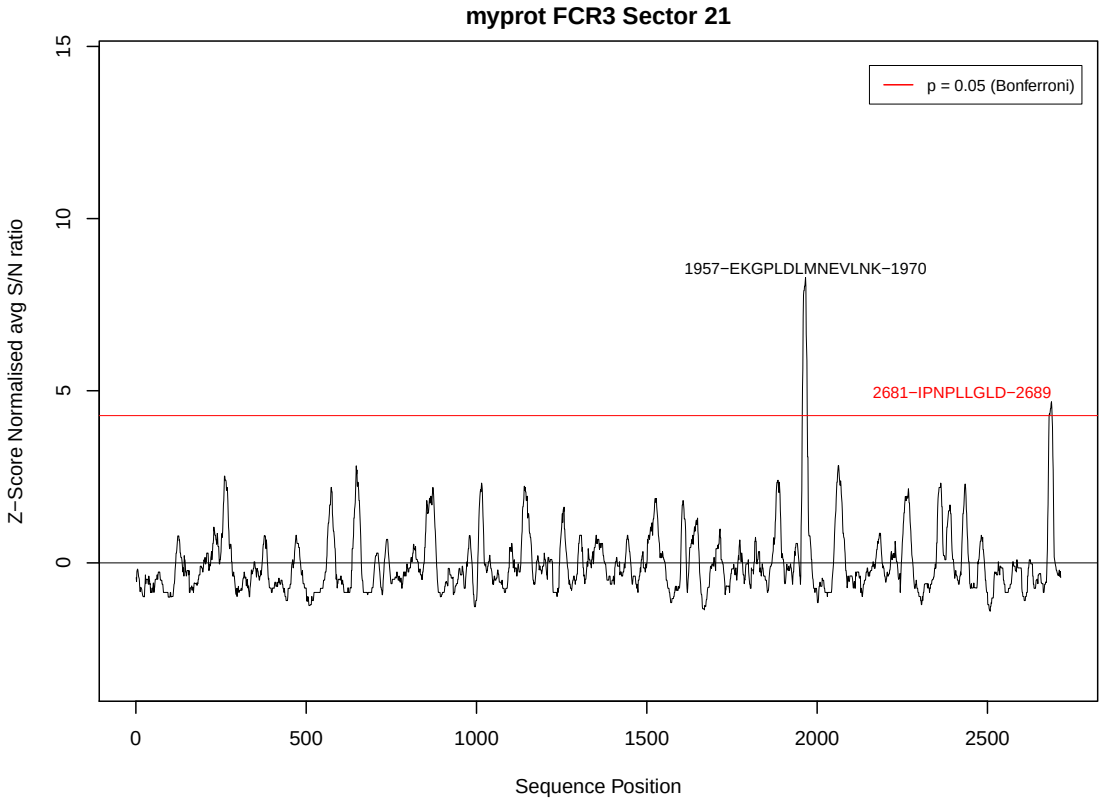

myprot FCR3 Sector 16
p = 0.05 (Bonferroni)
1938−TTSGTVNKKLQKKETEC−1954
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

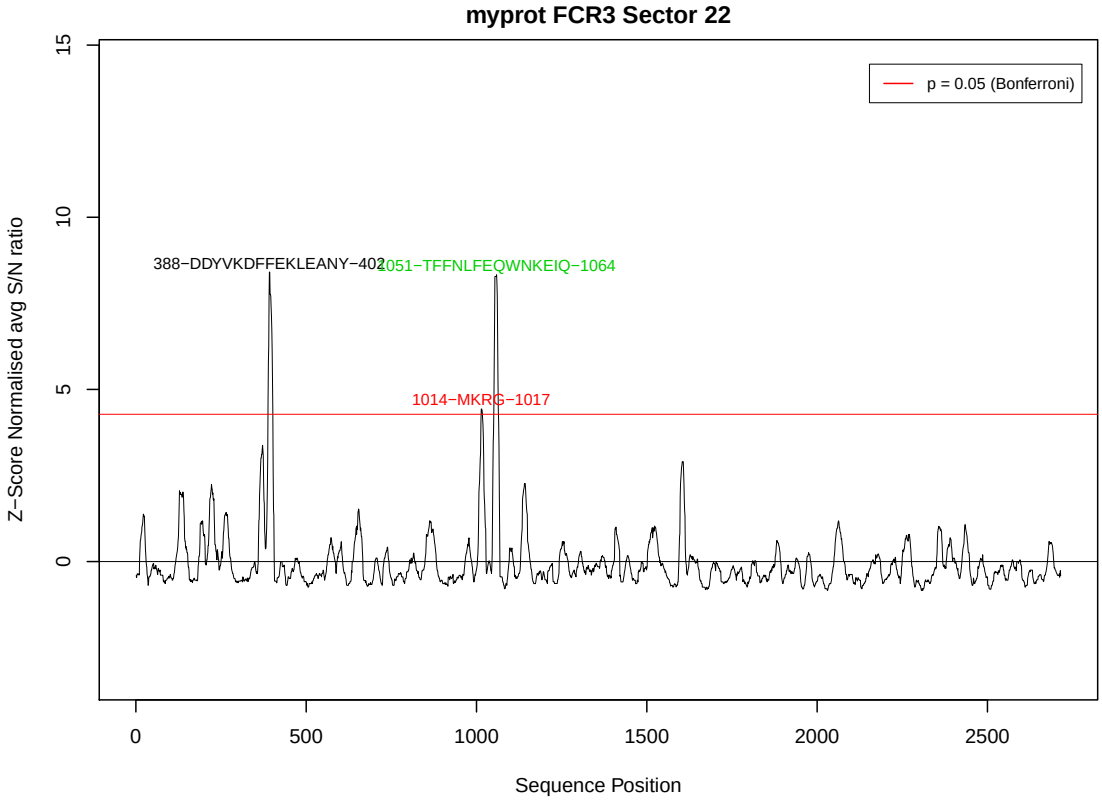

myprot FCR3 Sector 17
p = 0.05 (Bonferroni)
1954−CEEEKGPLDLMNEVL−1968
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

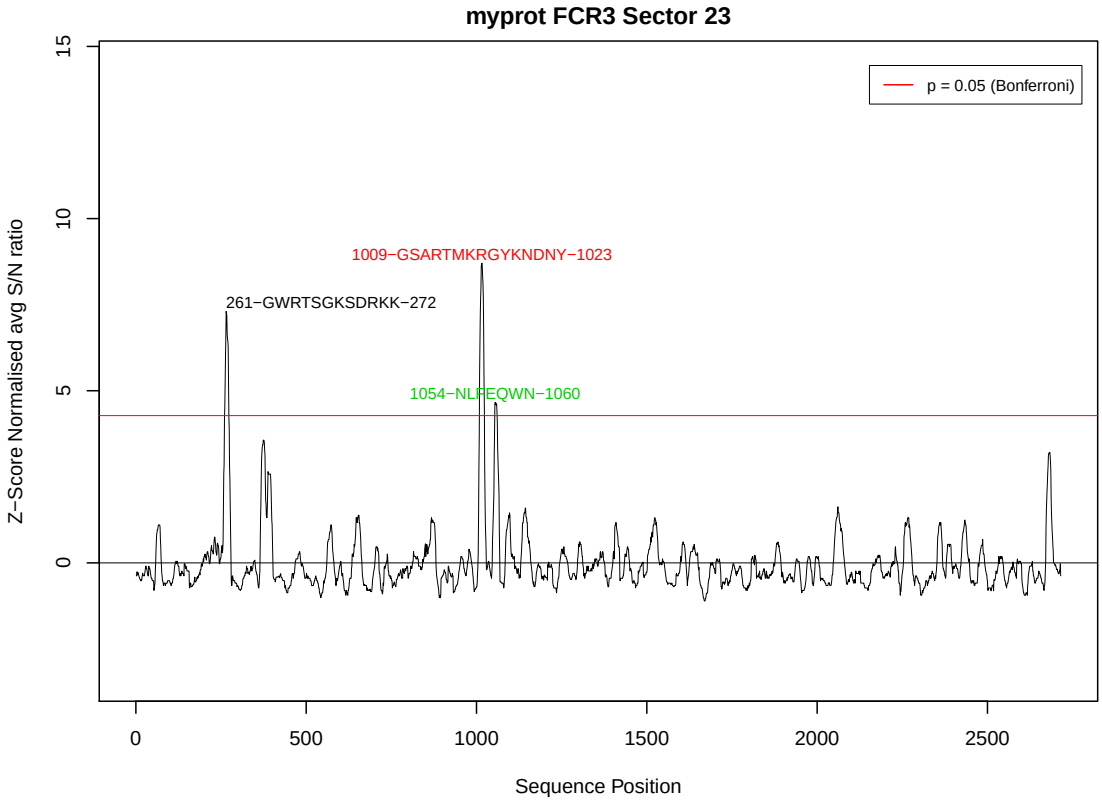

myprot FCR3 Sector 18
p = 0.05 (Bonferroni)
651−KRYPQNKNSGNKEN−664
1958−KGPLDLMNEVLNK−1970
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

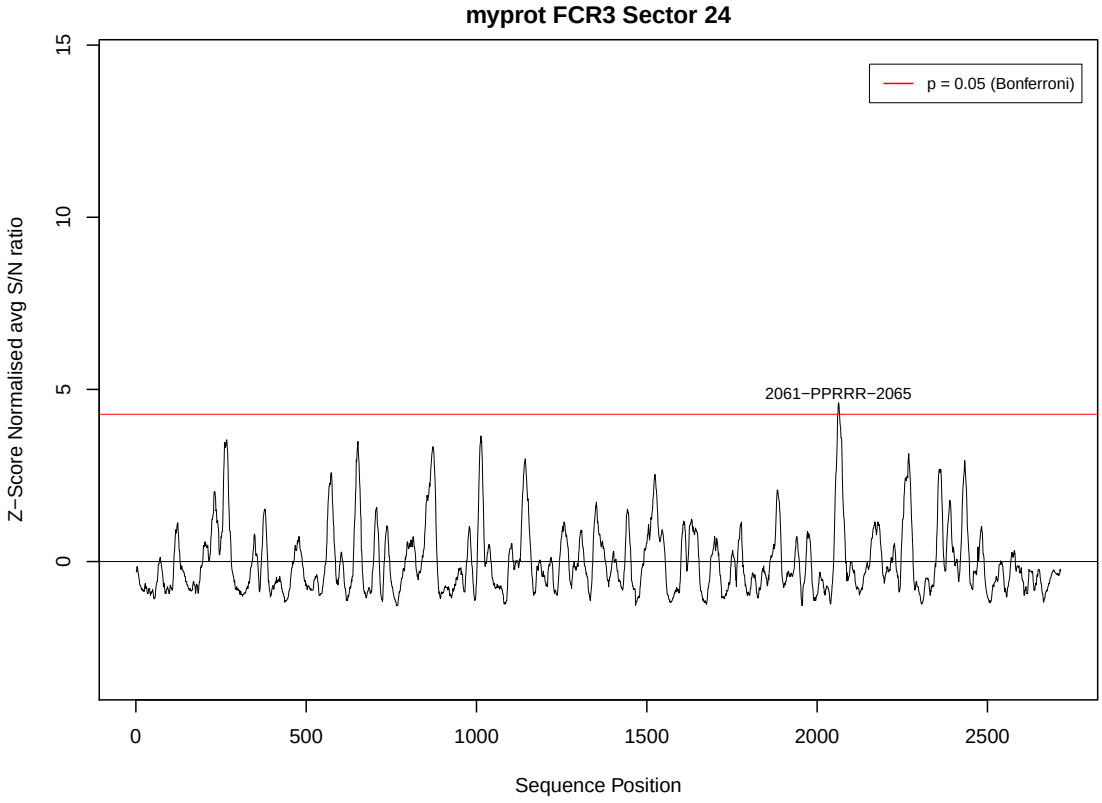

myprot FCR3 Sector 19
p = 0.05 (Bonferroni)
2681-I-2681
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
5
10
15


myprot FCR3 Sector 20
p = 0.05 (Bonferroni)
388−DDYVKDFFEKLEANYS−403
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
5
10
15


myprot FCR3 Sector 21
p = 0.05 (Bonferroni)
1957−EKGPLDLMNEVLNK−1970
2681−IPNPLLGLD−2689
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

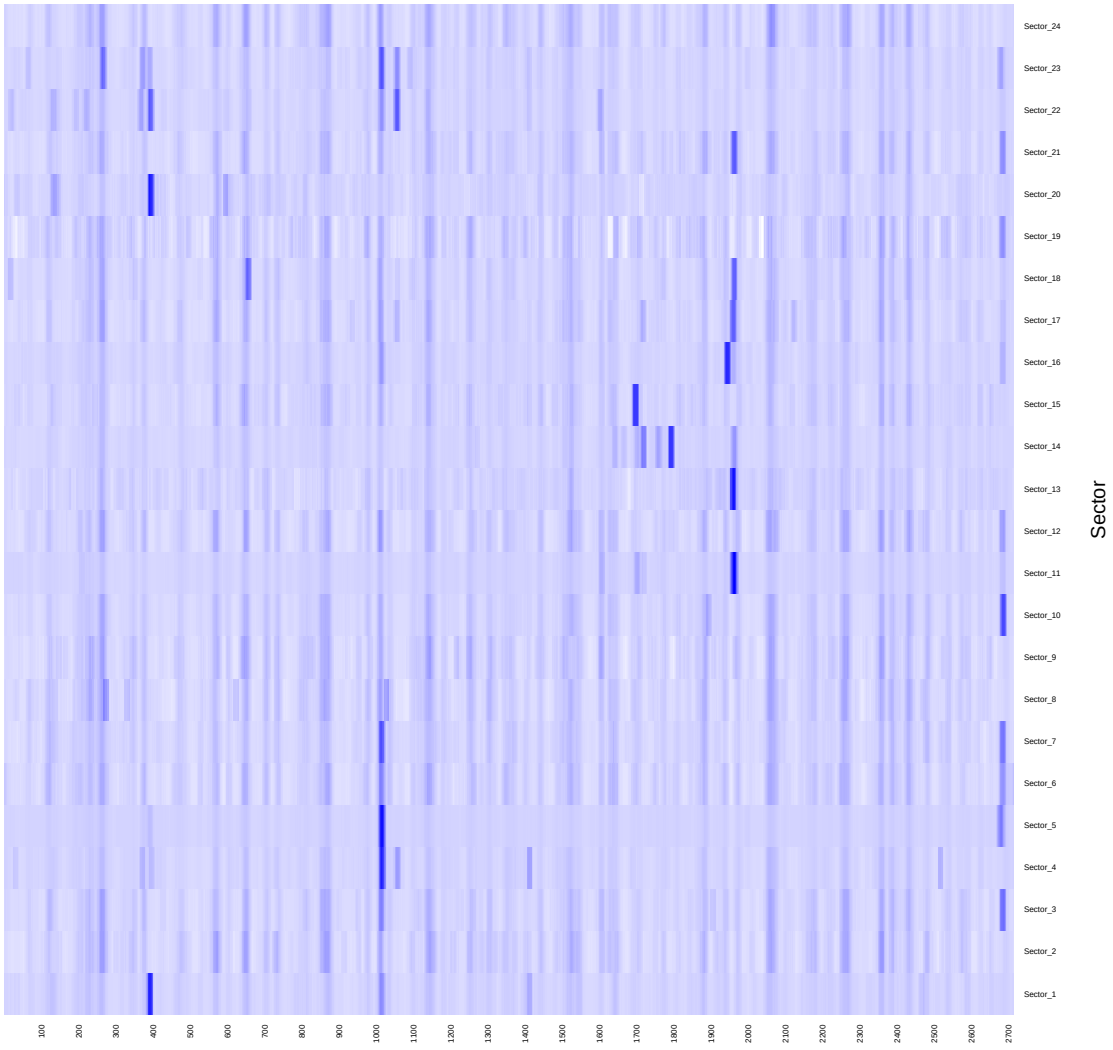

myprot FCR3 Sector 22
p = 0.05 (Bonferroni)
388−DDYVKDFFEKLEANY−402
2051−TFFNLFEQWNKEIQ−1064
1014−MKRG−1017
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500


myprot FCR3 Sector 23
p = 0.05 (Bonferroni)
1009−GSARTMKRGYKNDNY−1023
261−GWRTSGKSDRKK−272
1054−NLREQWN−1060
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

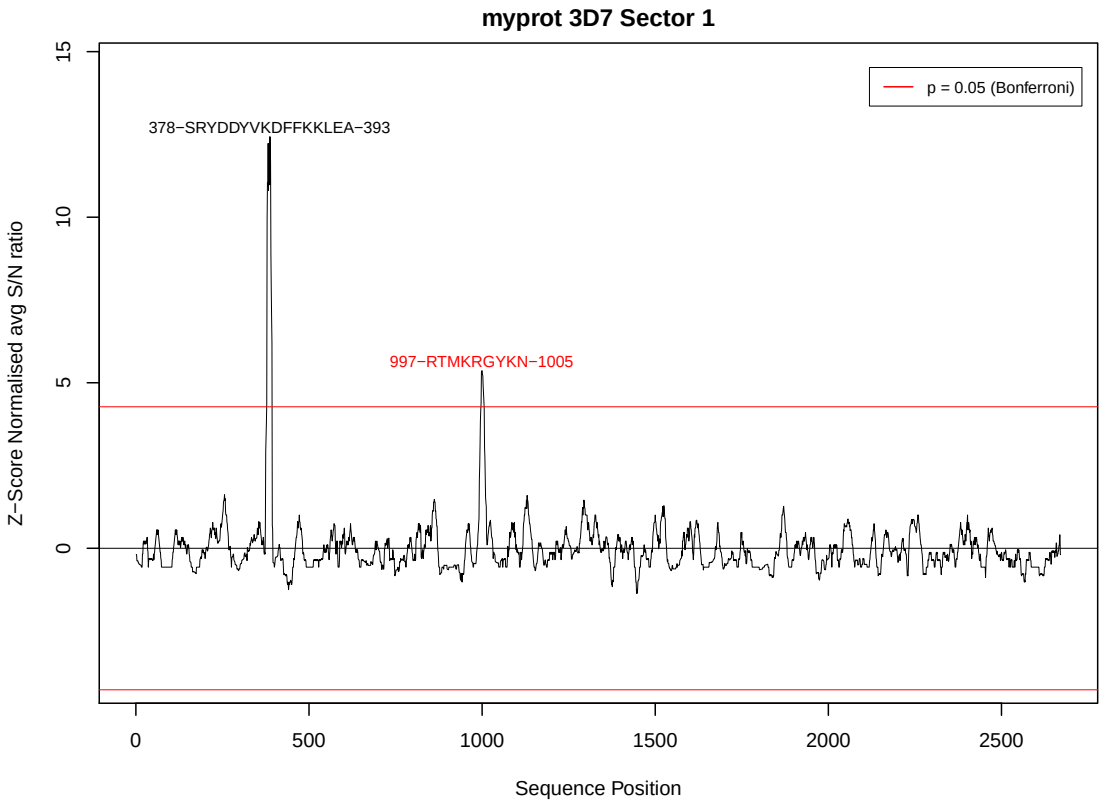

myprot FCR3 Sector 24
p = 0.05 (Bonferroni)
2061−PPRRR−2065
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

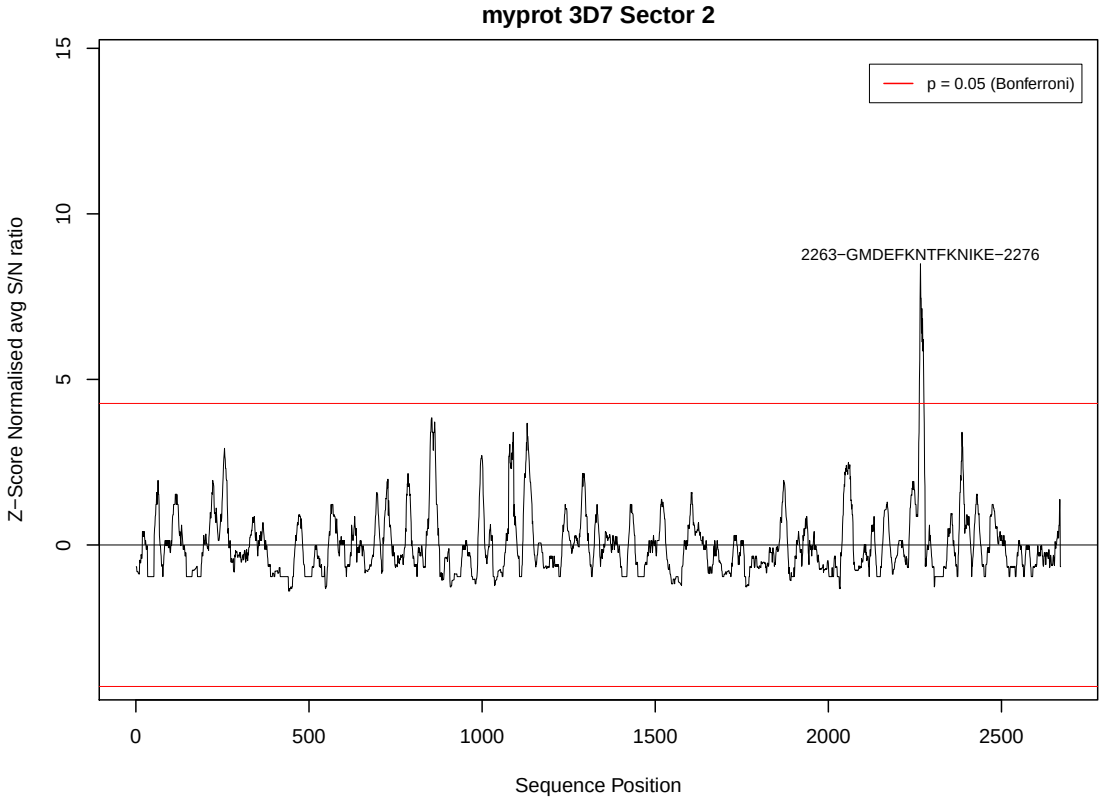
Heatmap of VAR2CSA FCR3 epitopes identified using the direct signal mapping approach for pan-sector signal identification

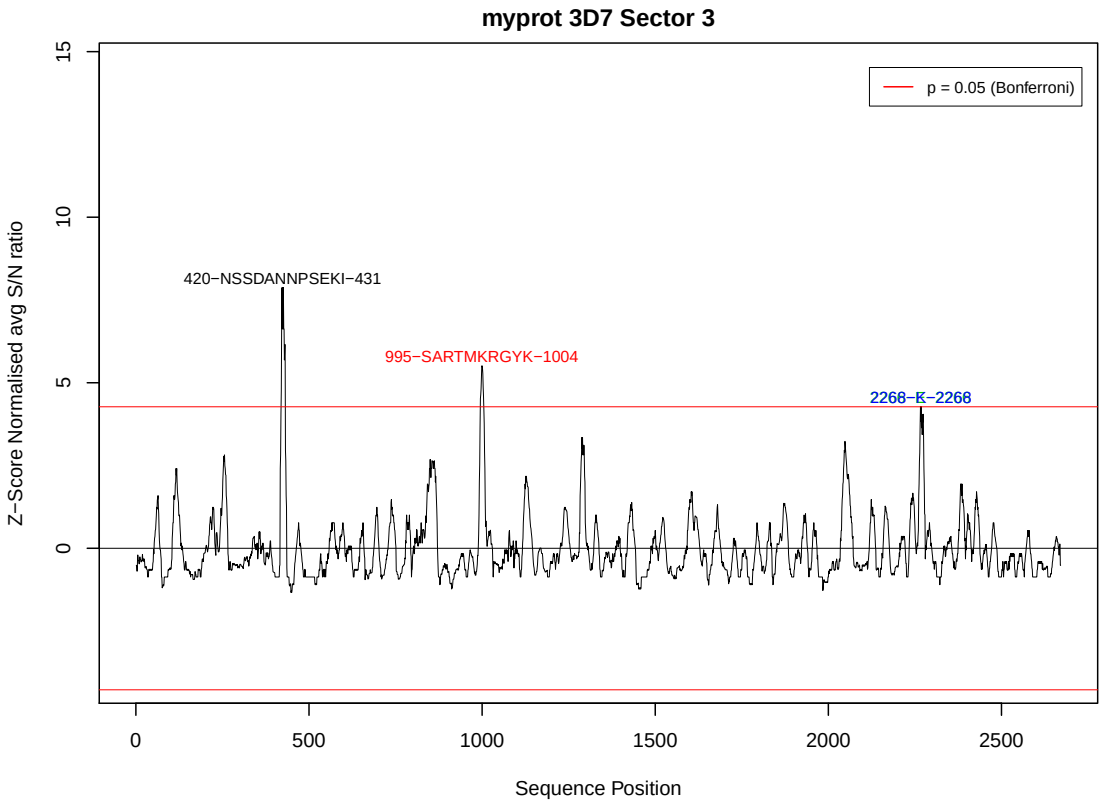
myprot FCR3

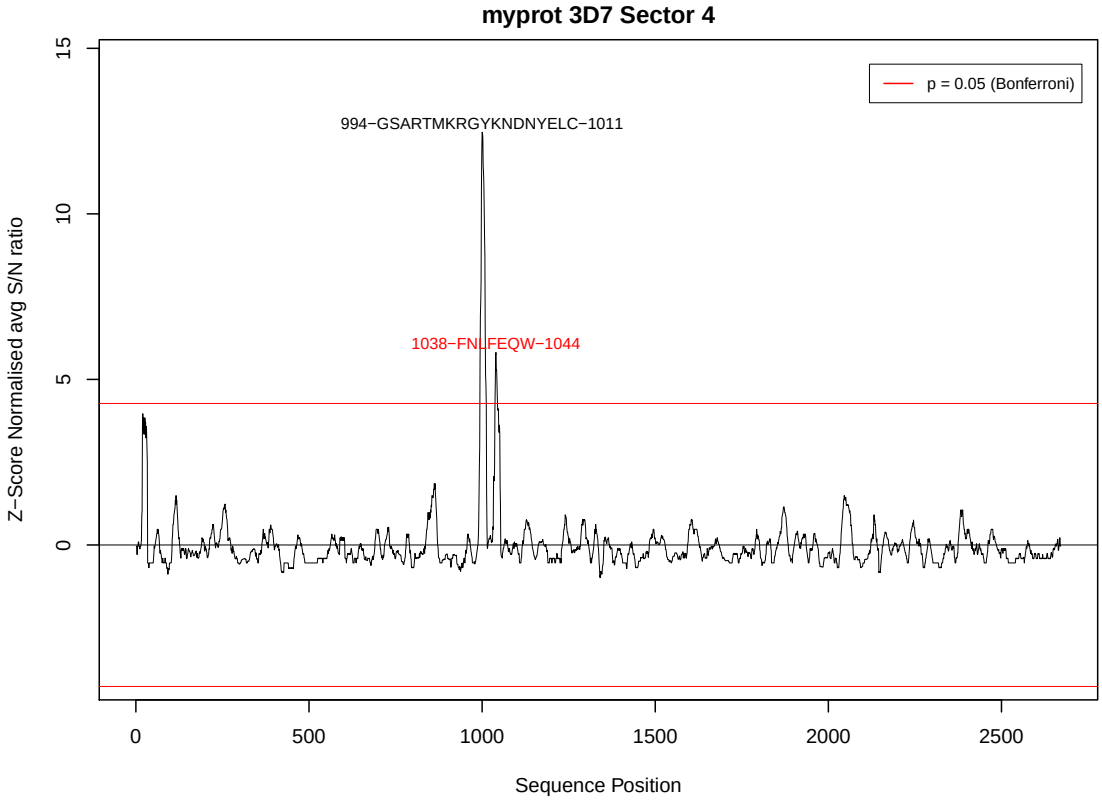

Sector_24
Sector_23
Sector_22
Sector_21
Sector_20
Sector_19
Sector_18
Sector_17
Sector_16
Sector_15
Sector_14
Sector_13
Sector_12
Sector_11
Sector_10
Sector_9
Sector_8
Sector_7
Sector_6
Sector_5
Sector_4
Sector_3
Sector_2
Sector_1
Sector
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
Sequence Position

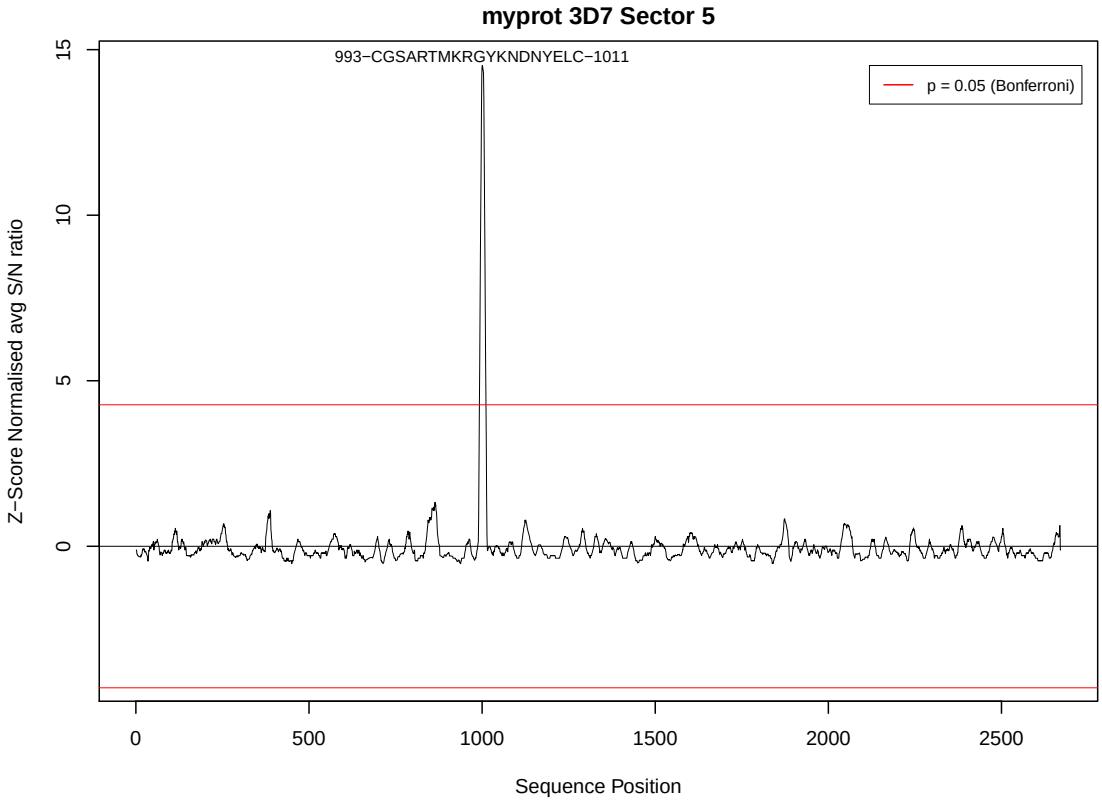
Plots of VAR2CSA 3D7 epitopes identified using the direct signal mapping approach

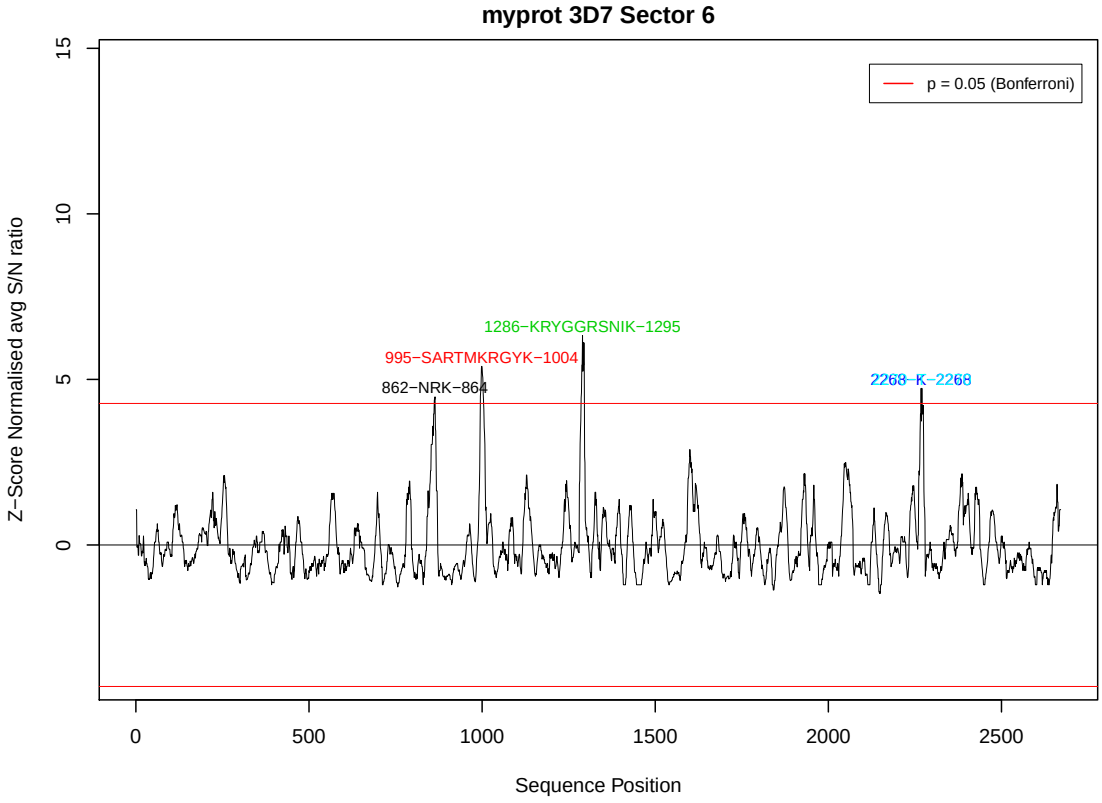

myprot 3D7 Sector 1
p = 0.05 (Bonferroni)
378−SRYDDYVKDFFKKLEA−393
997−RTMKRGYKN−1005
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

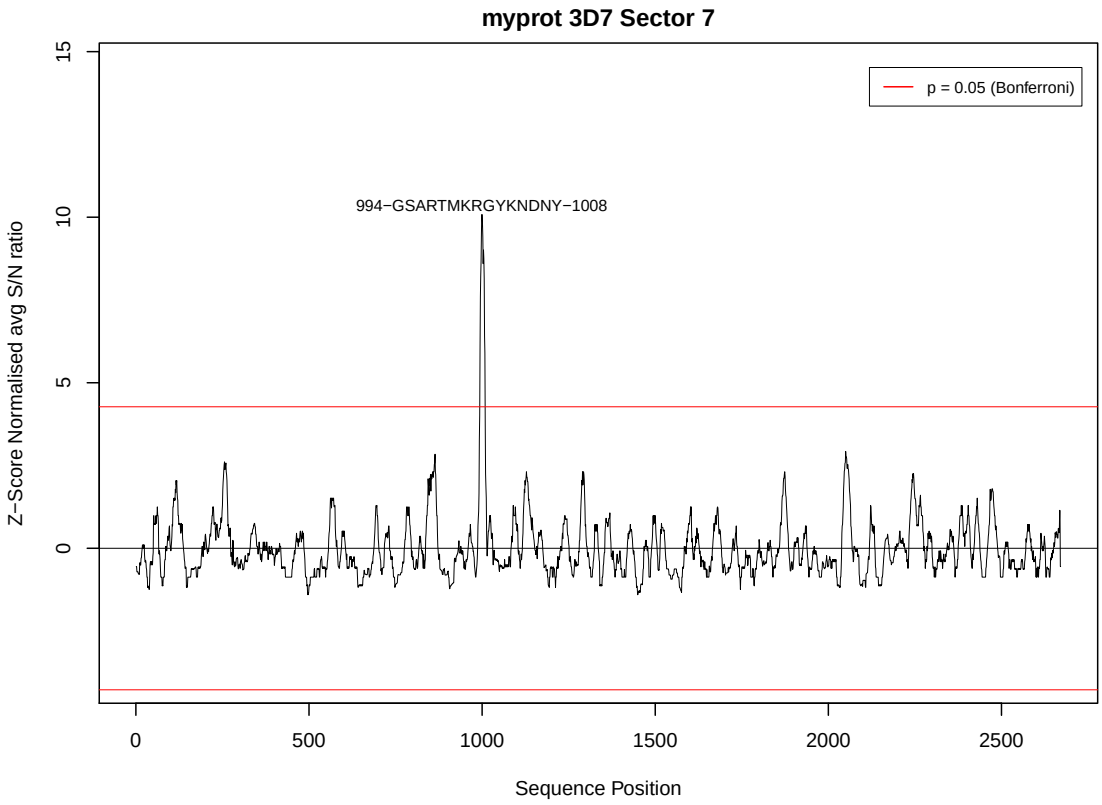

myprot 3D7 Sector 2
p = 0.05 (Bonferroni)
2263−GMDEFKNTFKNIKE−2276
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500

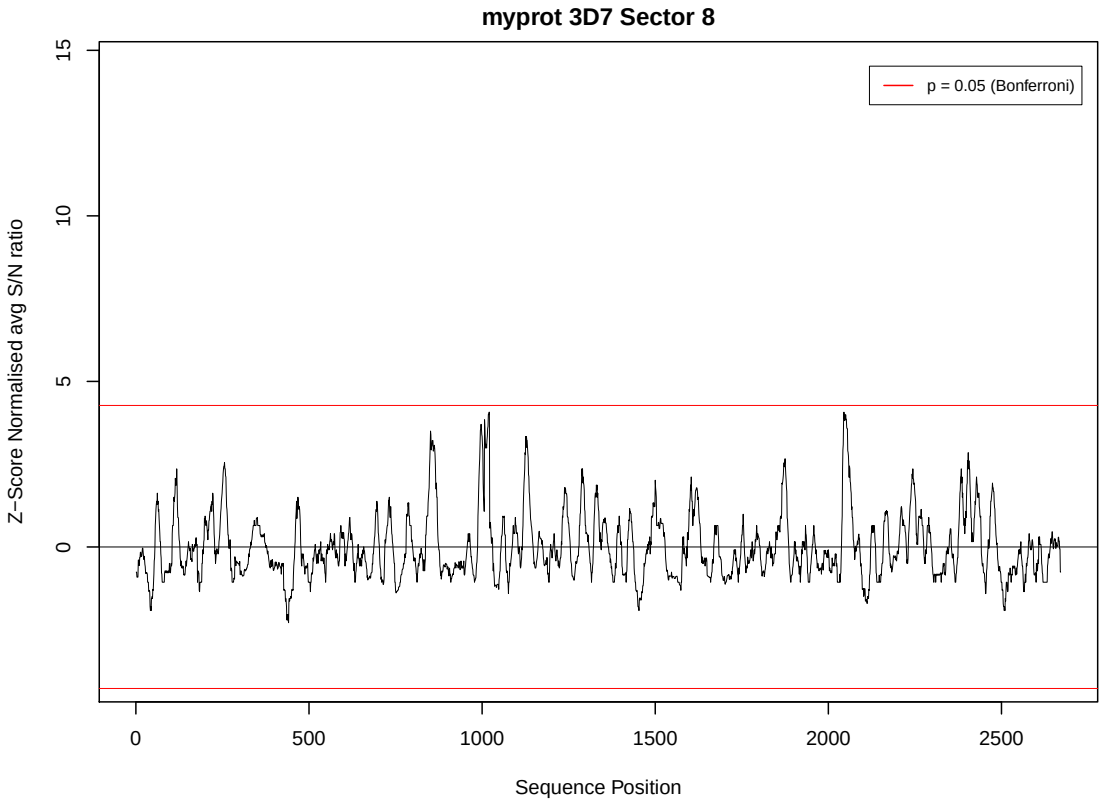

myprot 3D7 Sector 3
p = 0.05 (Bonferroni)
420−NSSDANNPSEKI−431
995−SARTMKRGYK−1004
2268−K−2268
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

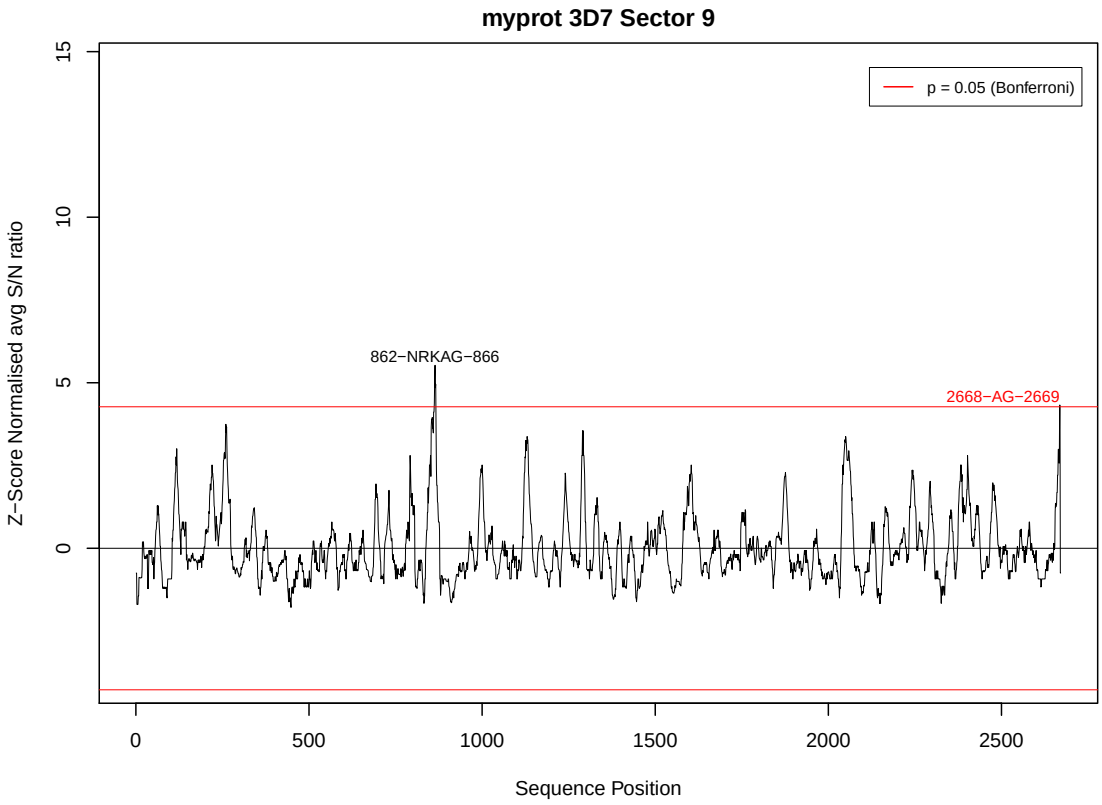

myprot 3D7 Sector 4
p = 0.05 (Bonferroni)
994−GSARTMKRGYKNDNYELC−1011
1038−FNLFEQW−1044
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

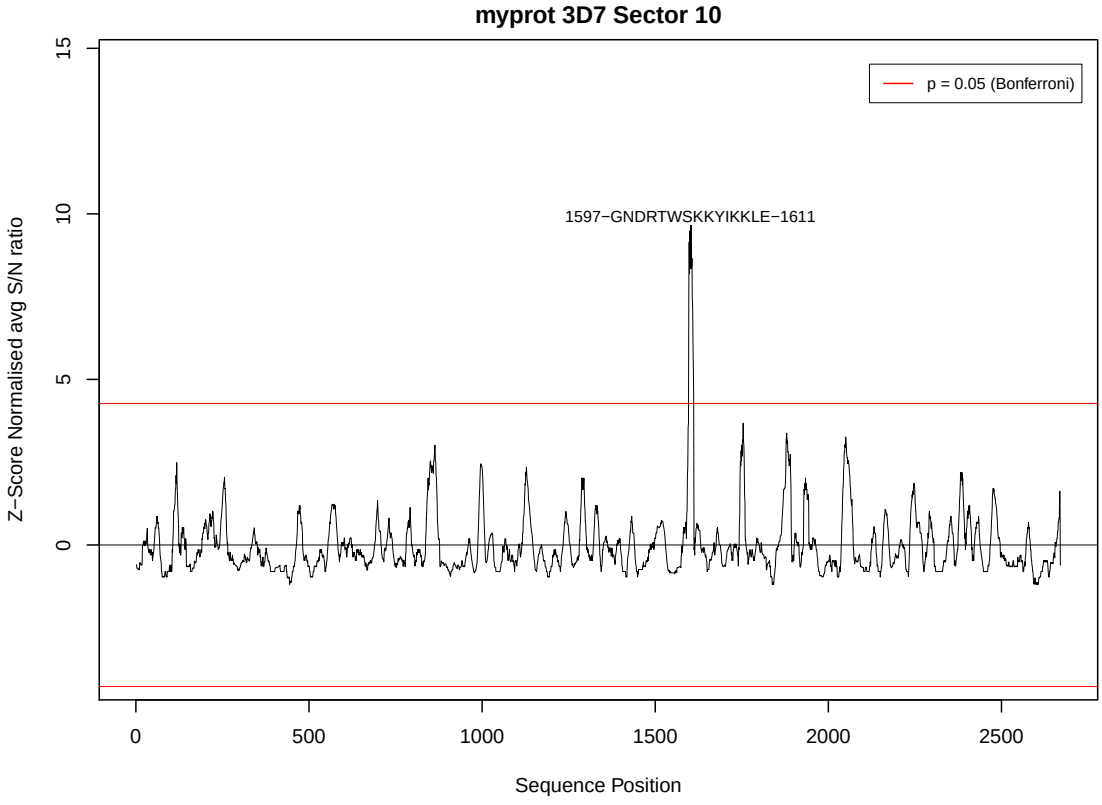

myprot 3D7 Sector 5
993−CGSARTMKRGYKNDNYELC−1011
p = 0.05 (Bonferroni)
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

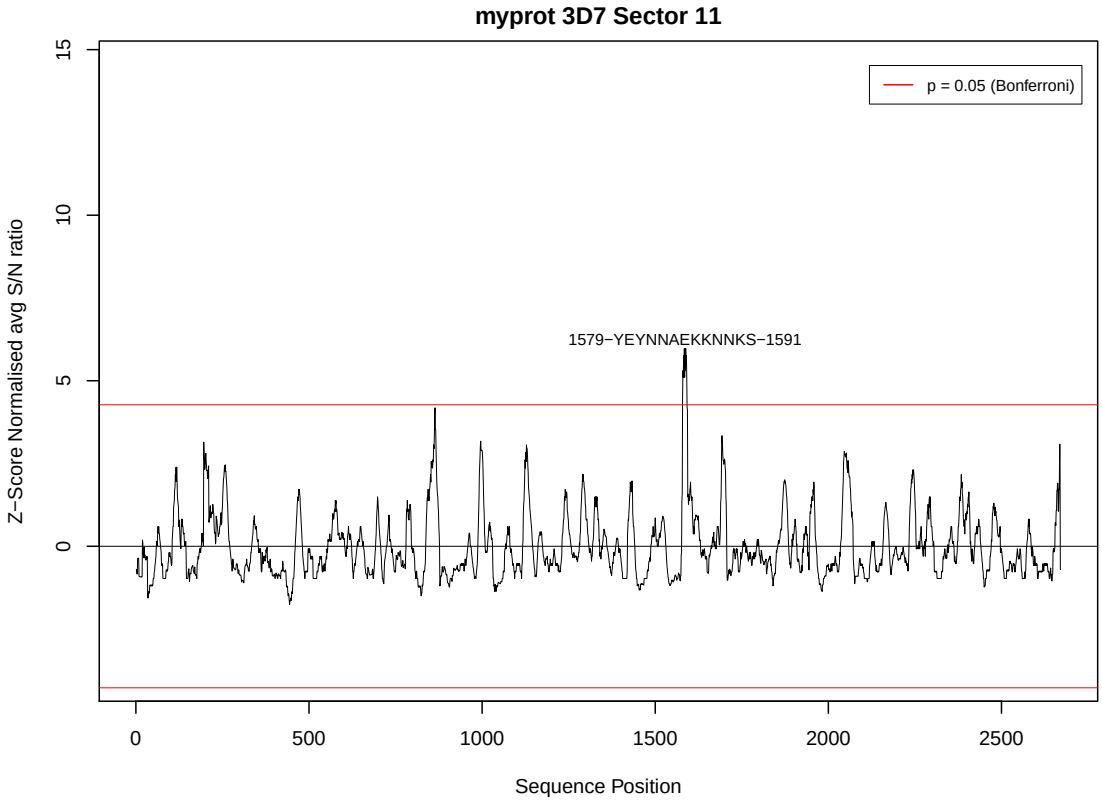

myprot 3D7 Sector 6
p = 0.05 (Bonferroni)
1286−KRYGGRSNIK−1295
995−SARTMKRGYK−1004
862−NRK−864
2268−K−2268
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

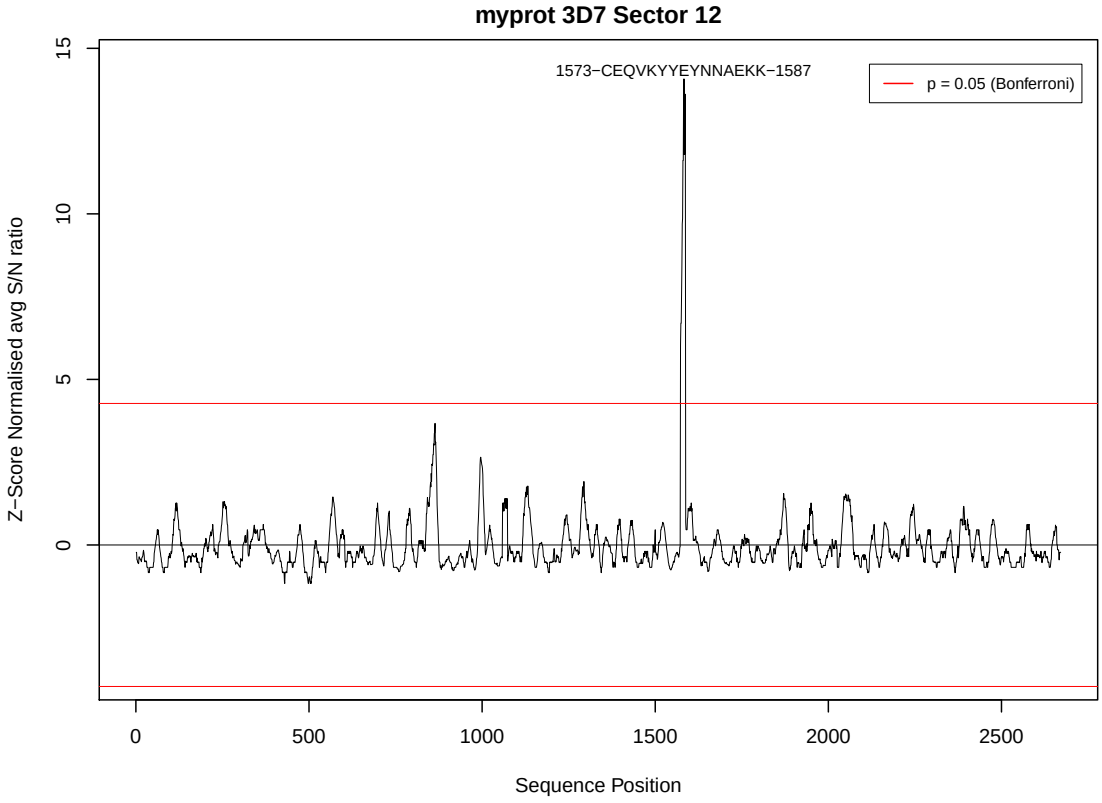

myprot 3D7 Sector 7
p = 0.05 (Bonferroni)
994−GSARTMKRGYKNDNY−1008
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
5
10
15

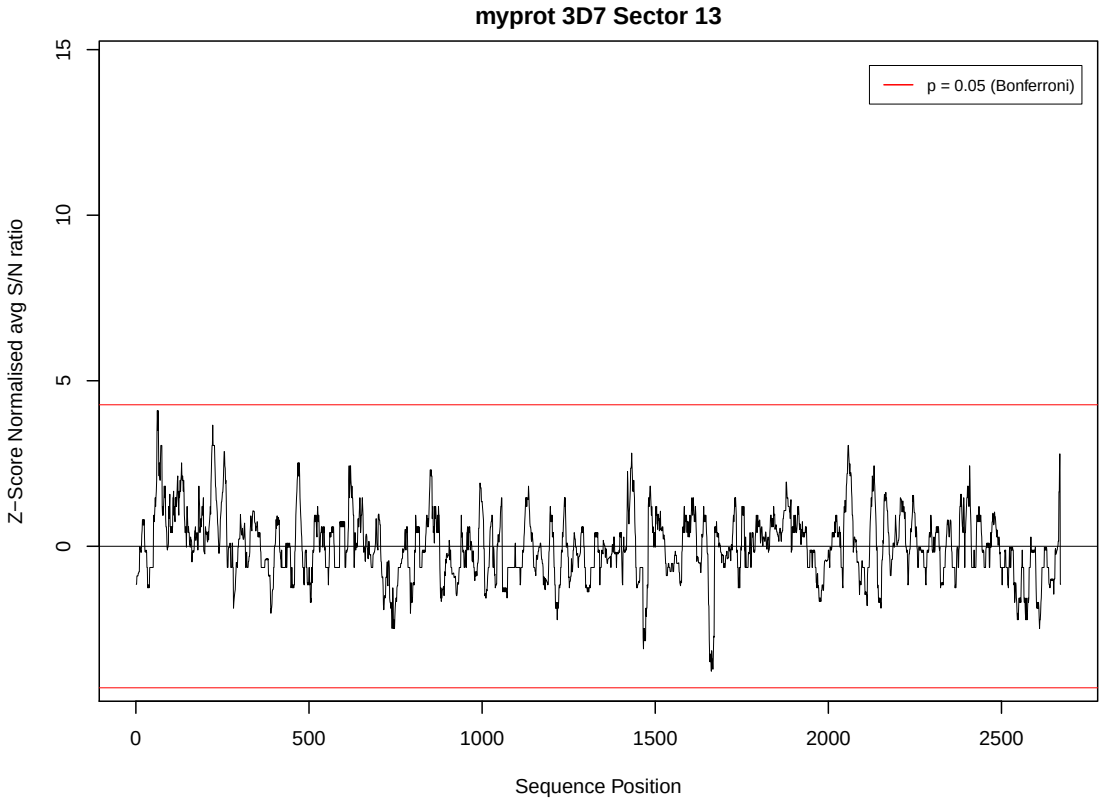

myprot 3D7 Sector 8
p = 0.05 (Bonferroni)
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
5
10
15

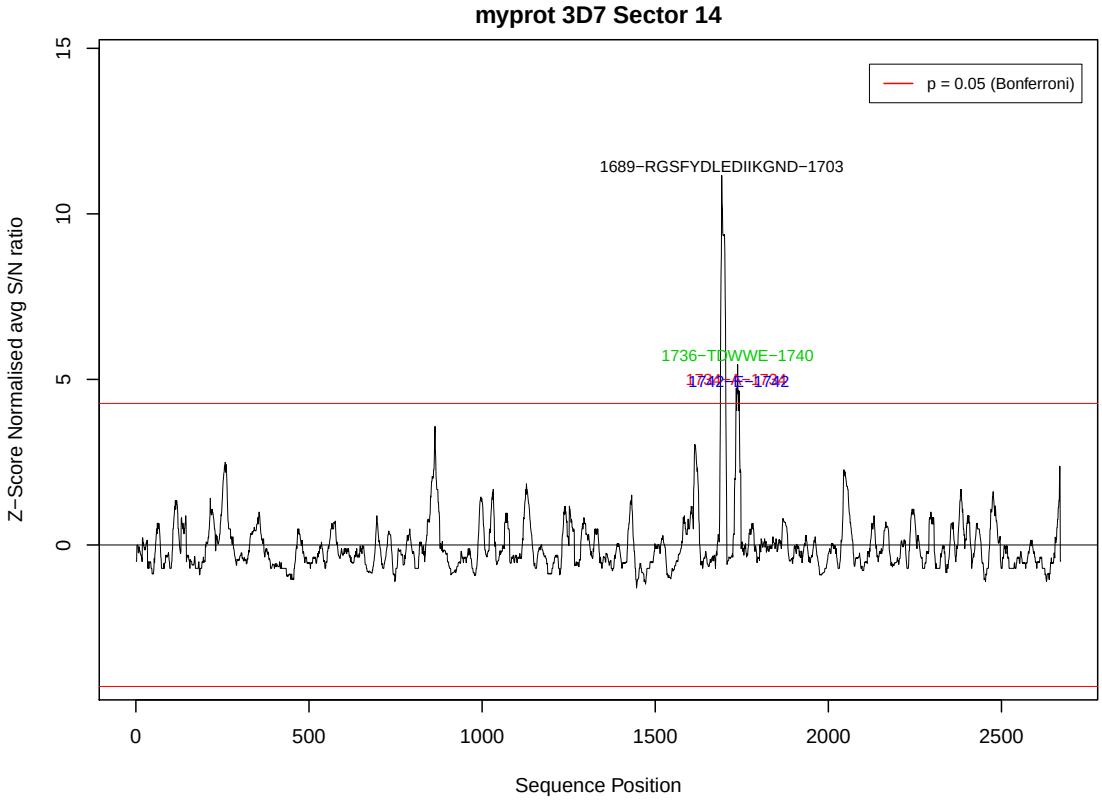

myprot 3D7 Sector 9
p = 0.05 (Bonferroni)
862−NRKAG−866
2668−AG−2669
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

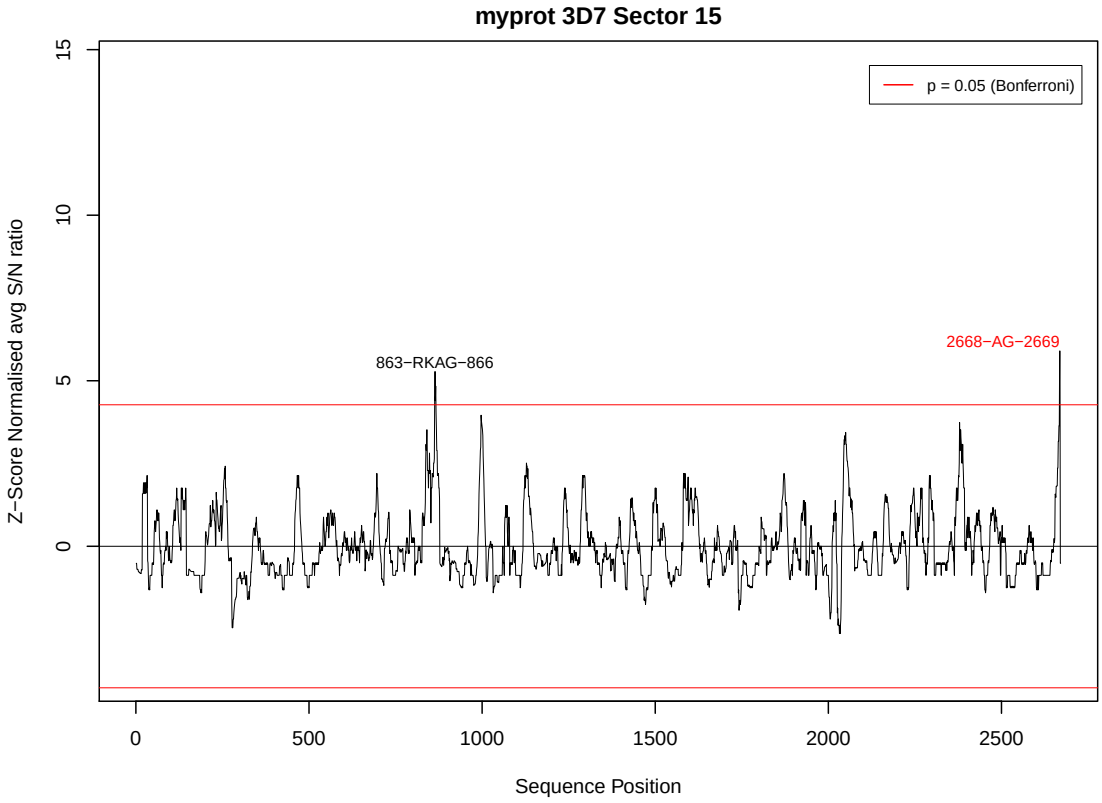

myprot 3D7 Sector 10
p = 0.05 (Bonferroni)
1597−GNDRTWSKKYIKKLE−1611
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

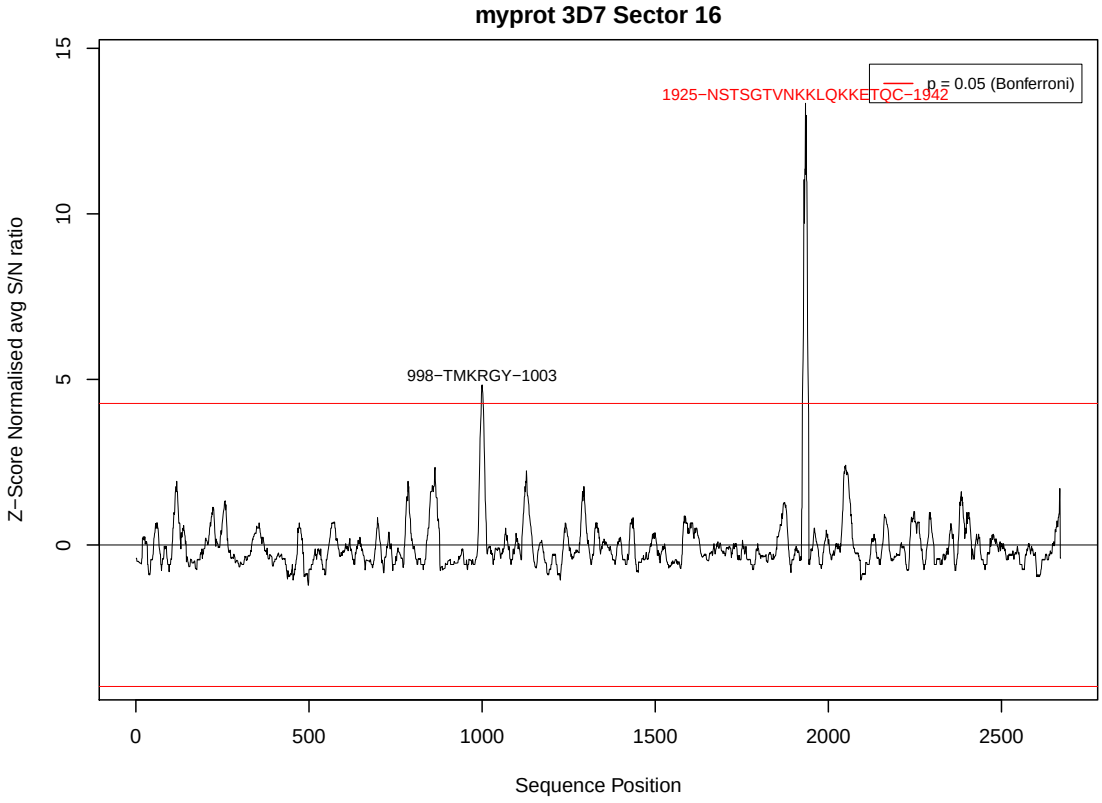

myprot 3D7 Sector 11
p = 0.05 (Bonferroni)
1579−YEYNNAEKKNNKS−1591
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

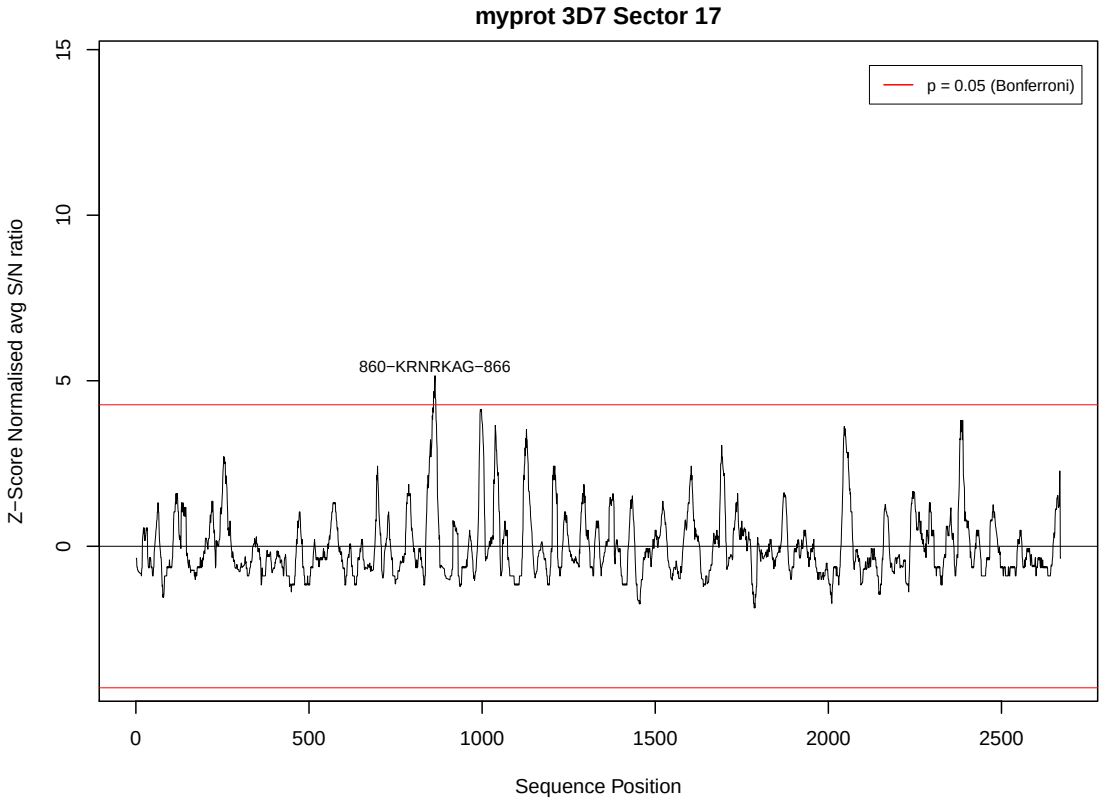

myprot 3D7 Sector 12
p = 0.05 (Bonferroni)
1573−CEQVKYYEYNNAEKK−1587
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

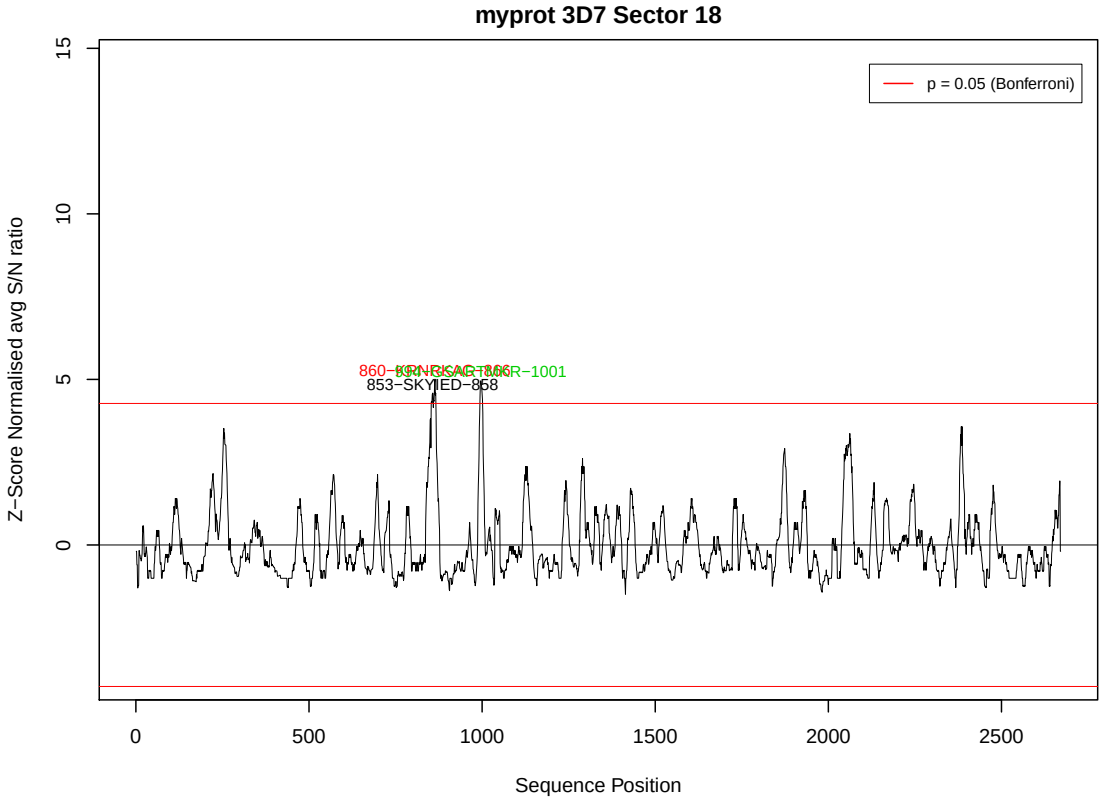

myprot 3D7 Sector 13
p = 0.05 (Bonferroni)
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

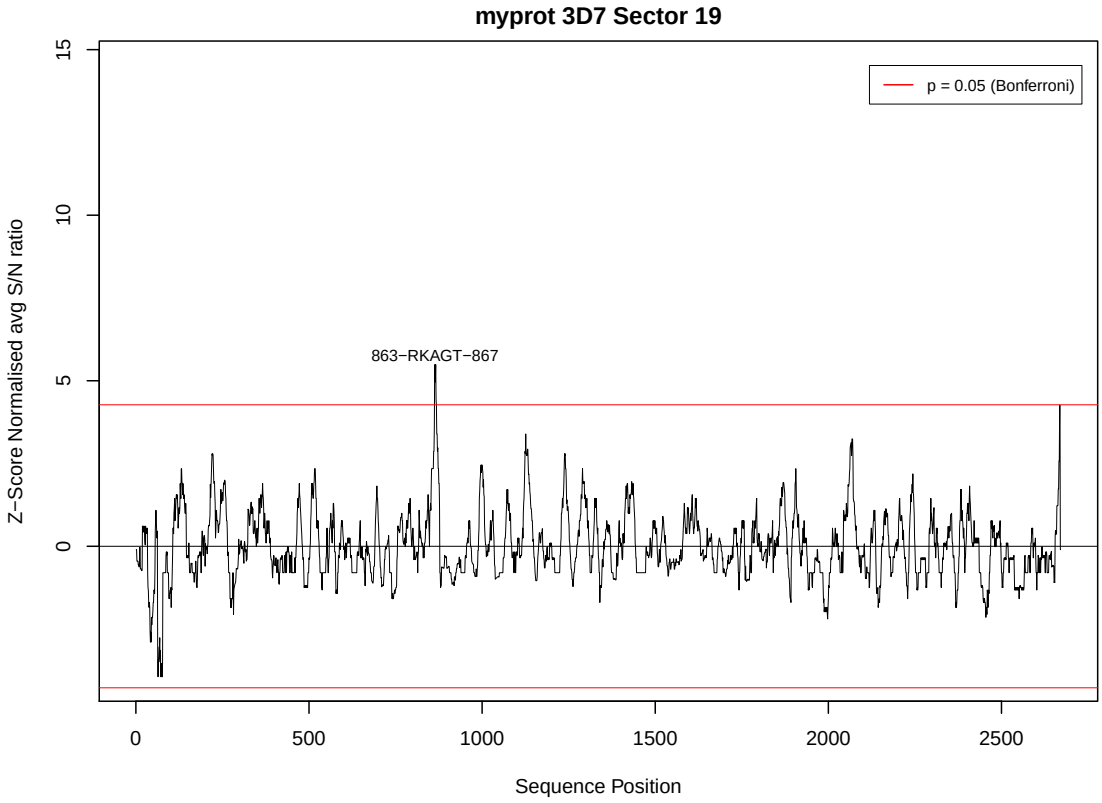

myprot 3D7 Sector 14
p = 0.05 (Bonferroni)
1689−RGSFYDLEDIIKGND−1703
1736−TDWWE−1740
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

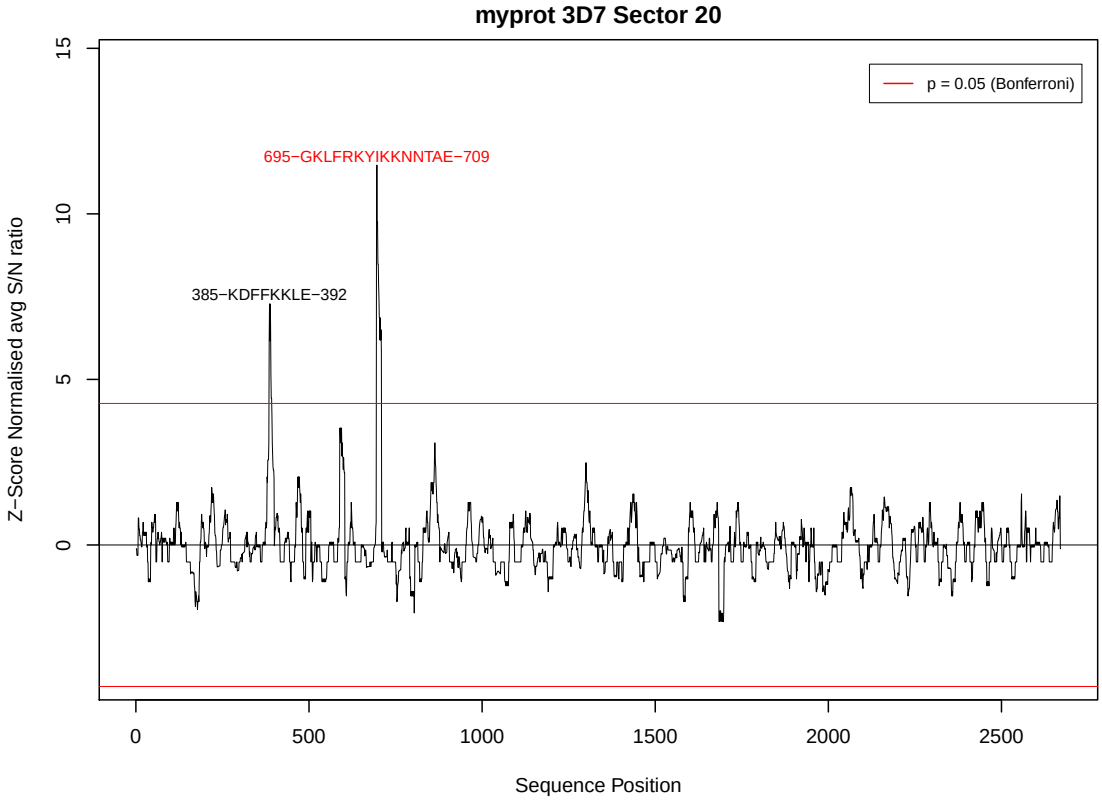

myprot 3D7 Sector 15
p = 0.05 (Bonferroni)
863−RKAG−866
2668−AG−2669
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

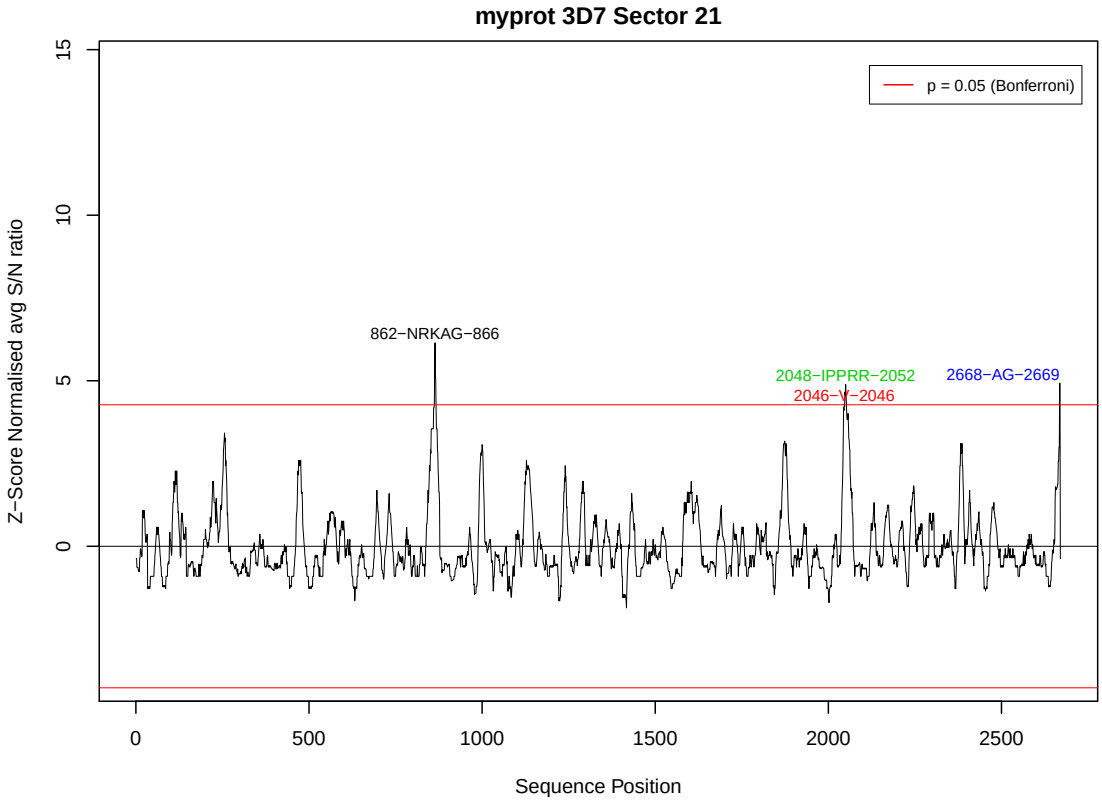

myprot 3D7 Sector 16
p = 0.05 (Bonferroni)
1925−NSTSGTVNKKLQKKETQC−1942
998−TMKRGY−1003
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

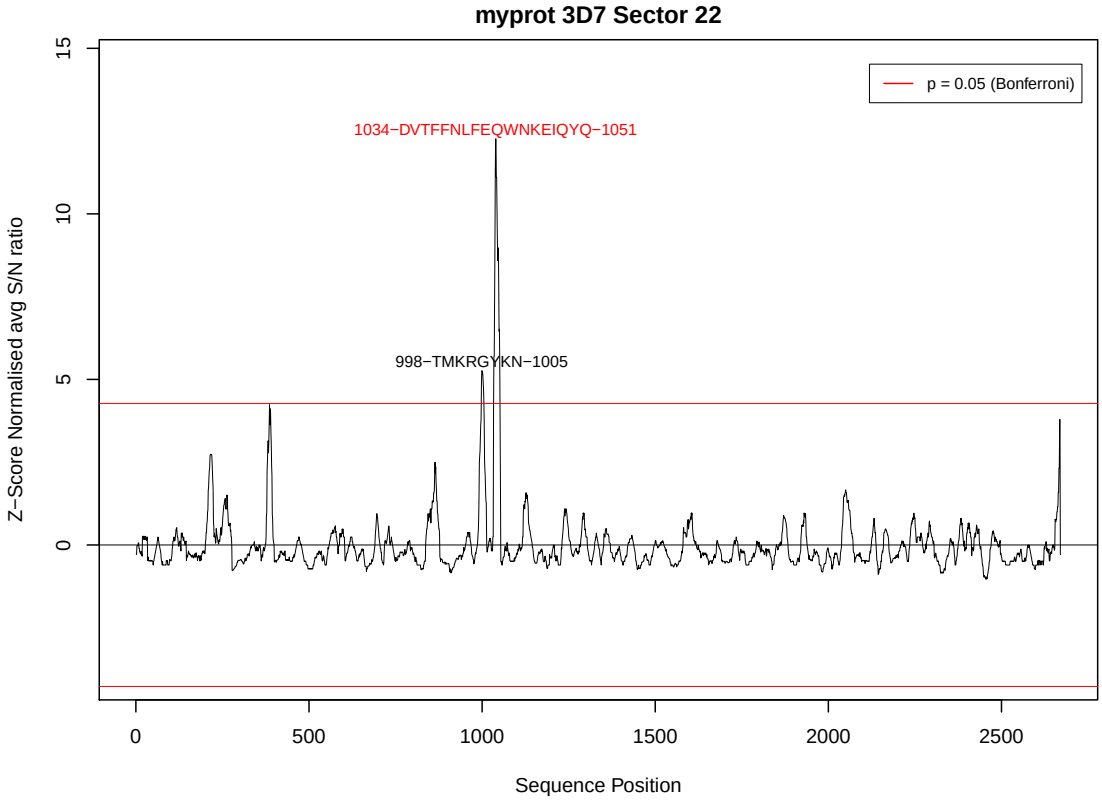

myprot 3D7 Sector 17
p = 0.05 (Bonferroni)
860−KRNRKAG−866
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

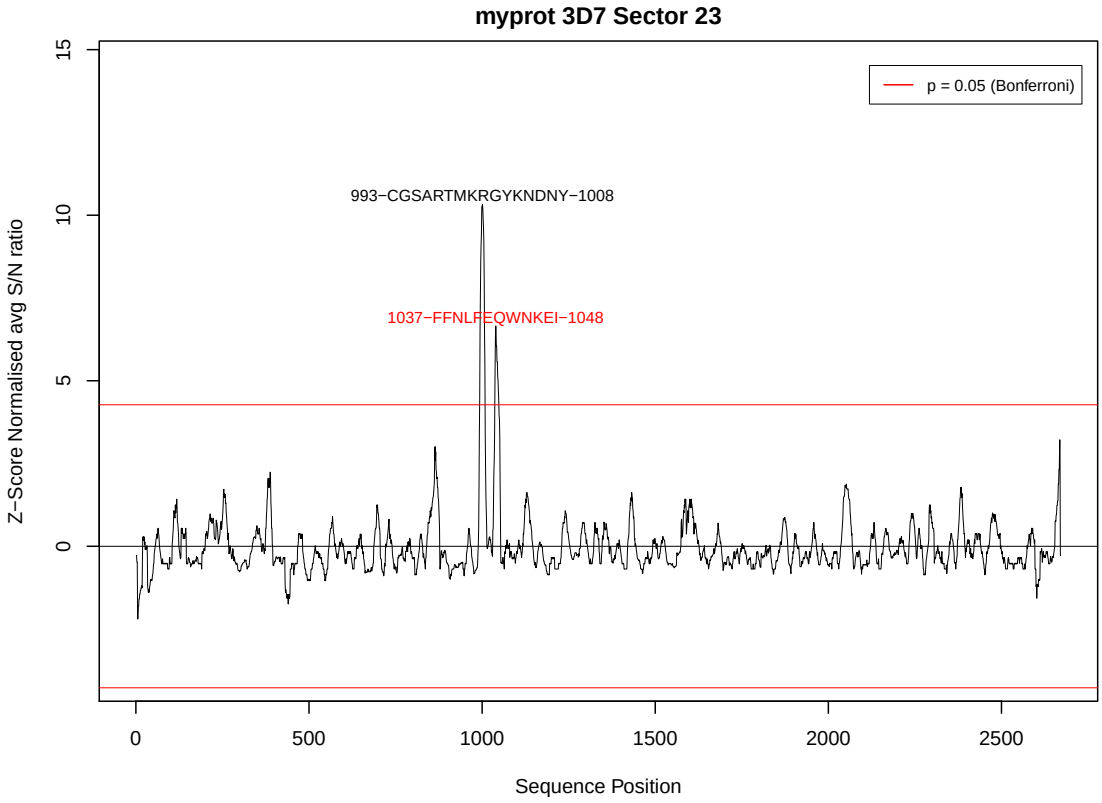

myprot 3D7 Sector 18
p = 0.05 (Bonferroni)
853−SKYIED−858
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

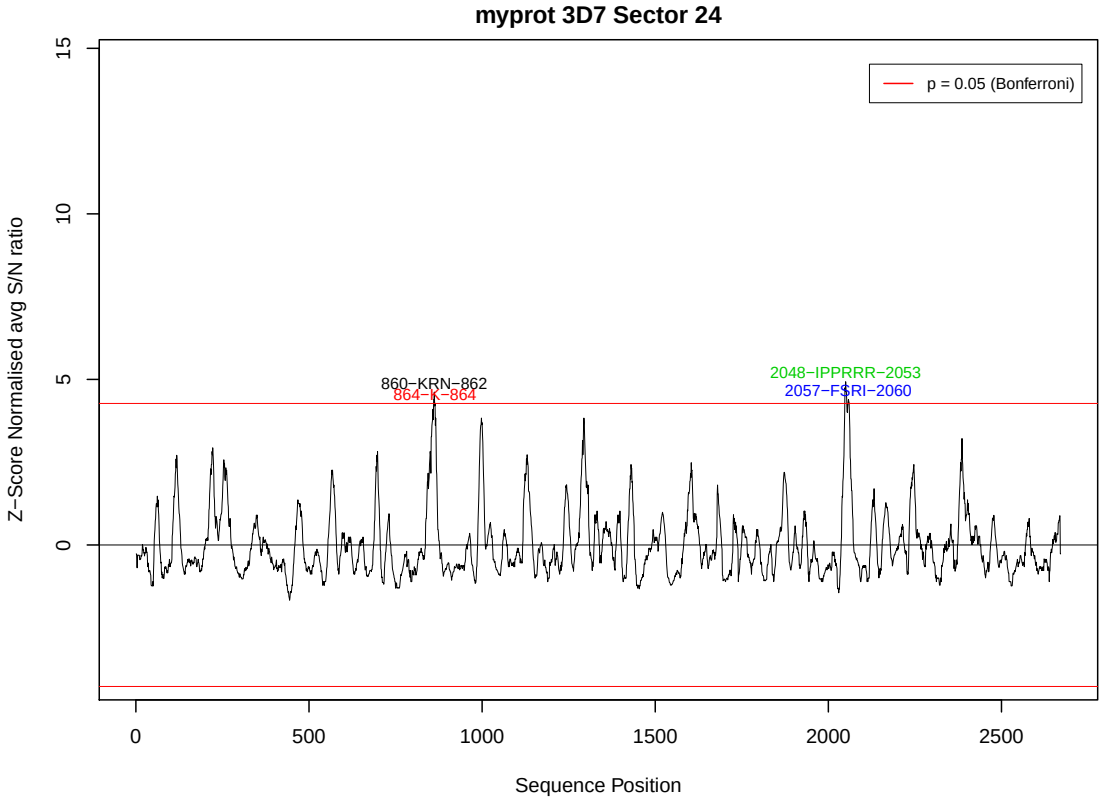

myprot 3D7 Sector 19
p = 0.05 (Bonferroni)
863−RKAGT−867
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500


myprot 3D7 Sector 20
p = 0.05 (Bonferroni)
695−GKLFRKYIKKNNTAE−709
385−KDFFKKLE−392
Z−Score Normalised avg S/N ratio
Sequence Position
0
5
10
15
0
500
1000
1500
2000
2500


myprot 3D7 Sector 21
p = 0.05 (Bonferroni)
862−NRKAG−866
2048−IPPRR−2052
2046−V−2046
2668−AG−2669
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

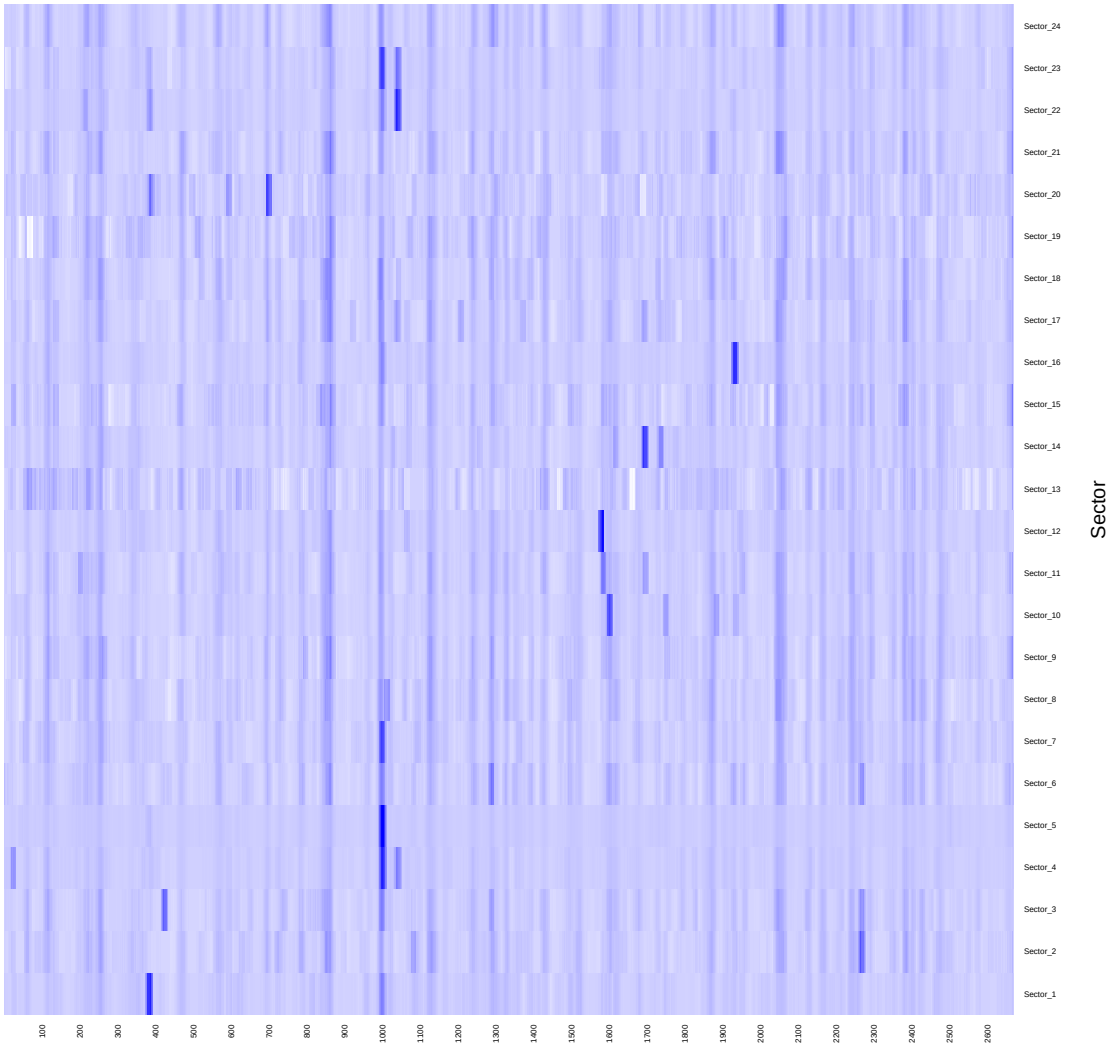

myprot 3D7 Sector 22
p = 0.05 (Bonferroni)
1034−DVTFFNLFEQWNKEIQYQ−1051
998−TMKRGYKN−1005
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15


myprot 3D7 Sector 23
p = 0.05 (Bonferroni)
993−CGSARTMKRGYKNDNY−1008
1037−FFNLFEQWNKEI−1048
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15


myprot 3D7 Sector 24
p = 0.05 (Bonferroni)
860−KRN−862
864−K−864
2048−IPPRRR−2053
2057−FSRI−2060
Z−Score Normalised avg S/N ratio
Sequence Position
0
500
1000
1500
2000
2500
0
5
10
15

Heatmap of VAR2CSA 3D7 epitopes identified using the direct signal mapping approach for pan-sector signal identification

myprot 3D7


Sector_24
Sector_23
Sector_22
Sector_21
Sector_20
Sector_19
Sector_18
Sector_17
Sector_16
Sector_15
Sector_14
Sector_13
Sector_12
Sector_11
Sector_10
Sector_9
Sector_8
Sector_7
Sector_6
Sector_5
Sector_4
Sector_3
Sector_2
Sector_1
Sector
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
Sequence Position

Tables of VAR2CSA 3D7 epitopes identified using the direct signal mapping approach

Protein	Sector	Start	End	Avg Z	Max Z	Sequence
myprot_3D7	1	378	393	9.994	12.425	SRYDDYVKDFFKKLEA
myprot_3D7	1	997	1005	4.99	5.363	RTMKRGYKN
myprot_3D7	2	2263	2276	6.401	8.491	GMDEFKNTFKNIKE
myprot_3D7	3	420	431	6.625	7.873	NSSDANNPSEKI
myprot_3D7	3	995	1004	5.035	5.513	SARTMKRGYK
myprot_3D7	3	2266	2266	4.283	4.283	E
myprot_3D7	3	2268	2268	4.283	4.283	K
myprot_3D7	4	994	1011	9.21	12.464	GSARTMKRGYKNDNYELC
myprot_3D7	4	1038	1044	5.17	5.815	FNLFEQW
myprot_3D7	5	993	1011	10.772	14.521	CGSARTMKRGYKNDNYELC
myprot_3D7	6	862	864	4.416	4.473	NRK
myprot_3D7	6	995	1004	4.927	5.396	SARTMKRGYK
myprot_3D7	6	1286	1295	5.456	6.323	KRYGGRSNIK
myprot_3D7	6	2268	2268	4.732	4.732	K
myprot_3D7	6	2270	2270	4.732	4.732	T
myprot_3D7	7	994	1008	8.419	10.082	GSARTMKRGYKNDNY
myprot_3D7	9	862	866	4.962	5.527	NRKAG
myprot_3D7	9	2668	2669	4.326	4.326	AG
myprot_3D7	10	1597	1611	7.906	9.651	GNDRTWSKKYIKKLE
myprot_3D7	11	1579	1591	5.422	5.977	YEYNNAEKKNNKS
myprot_3D7	12	1573	1587	10.34	14.072	CEQVKYYEYNNAEKK

Table 5.3.1: PepChipper-1.0 epitope list for VAR2CSA 3D7, sectors 1-12

Protein	Sector	Start	End	Avg Z	Max Z	Sequence
myprot_3D7	14	1689	1703	8.976	11.159	RGSFYDLEDIIKGND
myprot_3D7	14	1734	1734	4.739	4.739	A
myprot_3D7	14	1736	1740	4.83	5.448	TDWWE
myprot_3D7	14	1742	1742	4.654	4.654	E
myprot_3D7	15	863	866	4.729	5.279	RKAG
myprot_3D7	15	2668	2669	5.901	5.901	AG
myprot_3D7	16	998	1003	4.622	4.836	TMKRGY
myprot_3D7	16	1925	1942	8.88	13.341	NSTSGTVNKKLQKKETQC
myprot_3D7	17	860	866	4.683	5.151	KRNRKAG
myprot_3D7	18	853	858	4.441	4.585	SKYIED
myprot_3D7	18	860	866	4.617	5.007	KRNRKAG
myprot_3D7	18	994	1001	4.726	4.958	GSARTMKR
myprot_3D7	19	863	867	5.173	5.488	RKAGT
myprot_3D7	20	385	392	5.947	7.279	KDFFKKLE
myprot_3D7	20	695	709	7.893	11.475	GKLFRKYIKKNNTAE
myprot_3D7	21	862	866	5.212	6.145	NRKAG
myprot_3D7	21	2046	2046	4.289	4.289	V
myprot_3D7	21	2048	2052	4.604	4.883	IPPRR
myprot_3D7	21	2668	2669	4.929	4.929	AG
myprot_3D7	22	998	1005	4.968	5.263	TMKRGYKN
myprot_3D7	22	1034	1051	8.763	12.269	DVTFFNLFEQWNKEIQYQ
myprot_3D7	23	993	1008	8.453	10.325	CGSARTMKRGYKNDNY
myprot_3D7	23	1037	1048	5.55	6.653	FFNLFEQWNKEI
myprot_3D7	24	860	862	4.469	4.619	KRN
myprot_3D7	24	864	864	4.282	4.282	K
myprot_3D7	24	2048	2053	4.672	4.93	IPPRRR
myprot_3D7	24	2057	2060	4.327	4.394	FSRI

Table 5.3.2: PepChipper-1.0 epitope list for VAR2CSA 3D7, sectors 13-24

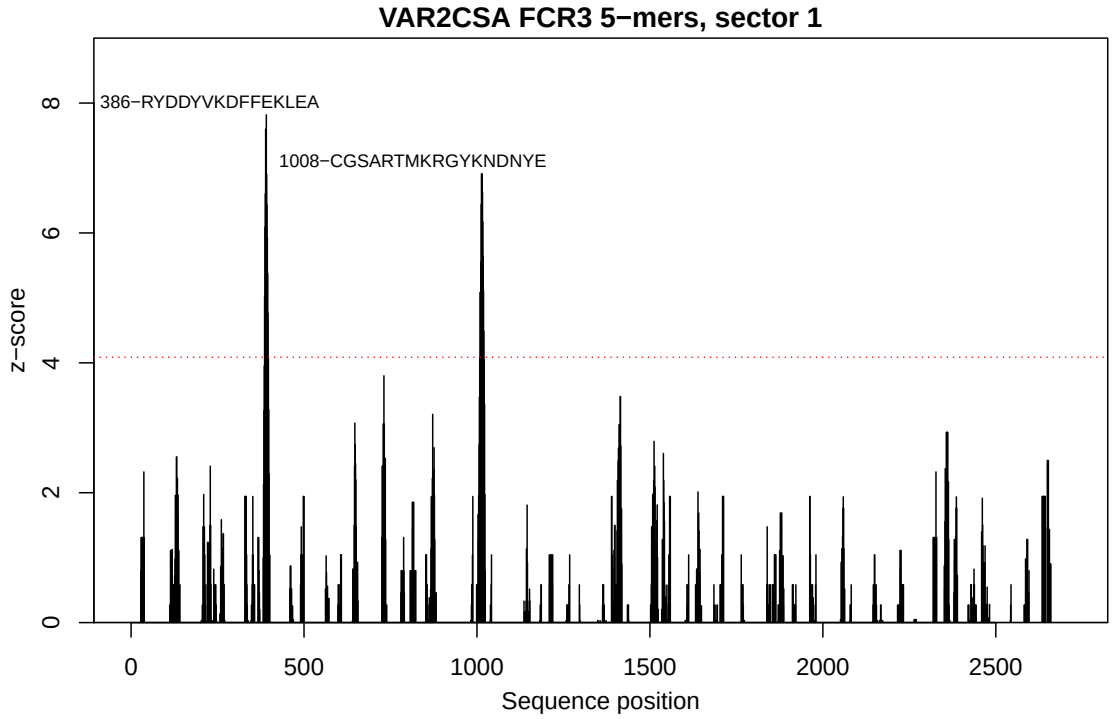
Tables of VAR2CSA FCR3 epitopes identified using the direct signal mapping approach

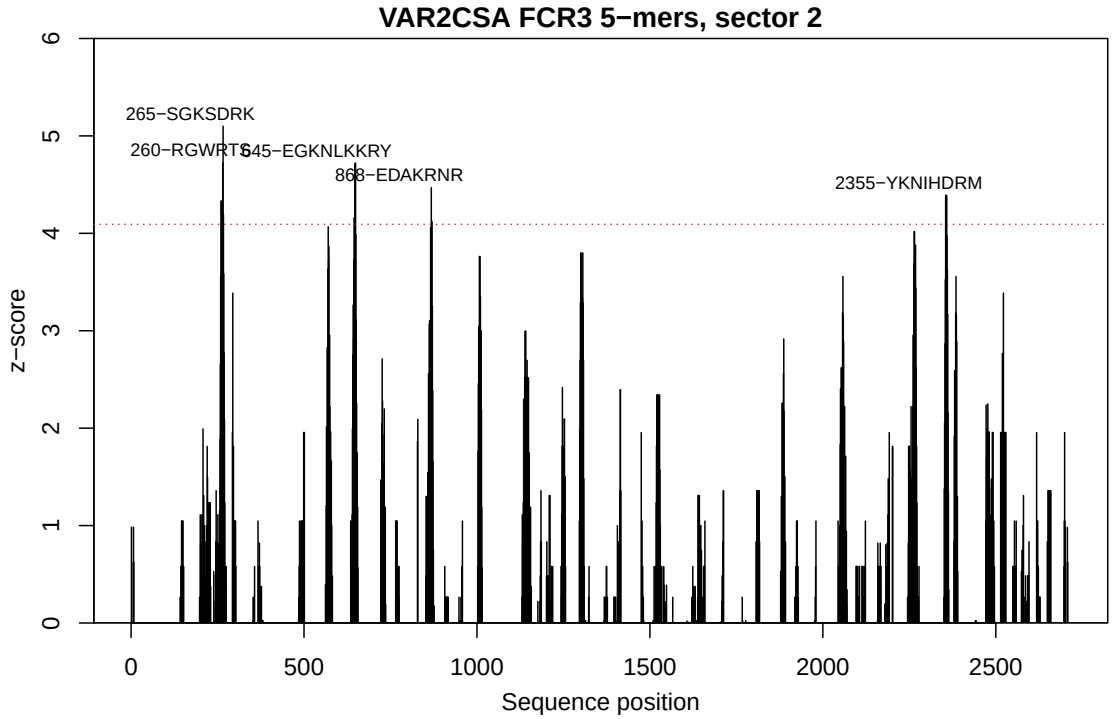
Protein	Sector	Start	End	Avg Z	Max Z	Sequence
myprot_FCR3	1	386	402	9.822	12.158	RYDDYVKDFFEKLEANY
myprot_FCR3	1	1013	1018	4.686	4.805	TMKRGY
myprot_FCR3	3	1010	1020	5.121	5.694	SARTMKRGYKN
myprot_FCR3	3	2680	2692	6.252	6.952	PIPNPLLGLDSTR
myprot_FCR3	4	1009	1026	9.182	12.45	GSARTMKRGYKNDNYELC
myprot_FCR3	5	1008	1025	10.212	13.349	CGSARTMKRGYKNDNYEL
myprot_FCR3	5	2676	2686	5.287	6.09	LEGKPIPNPLL
myprot_FCR3	6	1010	1020	5.115	5.683	SARTMKRGYKN
myprot_FCR3	7	1009	1023	7.747	9.287	GSARTMKRGYKNDNY
myprot_FCR3	7	2680	2692	5.798	6.383	PIPNPLLGLDSTR
myprot_FCR3	8	268	273	5.017	5.373	SDRKKN
myprot_FCR3	10	2680	2694	8.064	9.476	PIPNPLLGLDSTRTG
myprot_FCR3	11	1954	1973	9.872	14.446	CEEEKGPLDLMNEVLNKMDK
myprot_FCR3	13	1953	1967	11.162	13.205	ECEEEKGPLDLMNEV
myprot_FCR3	14	1715	1726	5.381	5.88	SFYDLEDIIKGN
myprot_FCR3	14	1787	1802	9.273	10.774	SGVRYAVEEKNENFPL
myprot_FCR3	15	1691	1706	9.285	10.534	WKQYNPTGKGIDDANK
myprot_FCR3	16	1938	1954	9.931	11.824	TTSGTVNKKLQKKETEC
myprot_FCR3	17	1954	1968	6.845	7.925	CEEEKGPLDLMNEVL
myprot_FCR3	18	651	664	6.313	7.618	KRYPQNKNSGNKEN
myprot_FCR3	18	1958	1970	7.09	7.842	KGPLDLMNEVLNK
myprot_FCR3	19	2681	2681	4.453	4.453	I
myprot_FCR3	19	2684	2688	4.445	4.531	PLLGL
myprot_FCR3	20	388	403	10.39	13.021	DDYVKDFFEKLEANYS
myprot_FCR3	21	1957	1970	7.198	8.29	EKGPLDLMNEVLNK
myprot_FCR3	21	2681	2689	4.455	4.682	IPNPLLGLD
myprot_FCR3	22	388	402	6.775	8.41	DDYVKDFFEKLEANY
myprot_FCR3	22	1014	1017	4.399	4.432	MKRG
myprot_FCR3	22	1051	1064	7.075	8.33	TFFNLFEQWNKEIQ
myprot_FCR3	23	261	272	6.254	7.303	GWRTSGKSDRKK
myprot_FCR3	23	1009	1023	7.329	8.704	GSARTMKRGYKNDNY
myprot_FCR3	23	1054	1060	4.601	4.656	NLFEQWN
myprot_FCR3	24	2061	2065	4.488	4.613	PPRRR

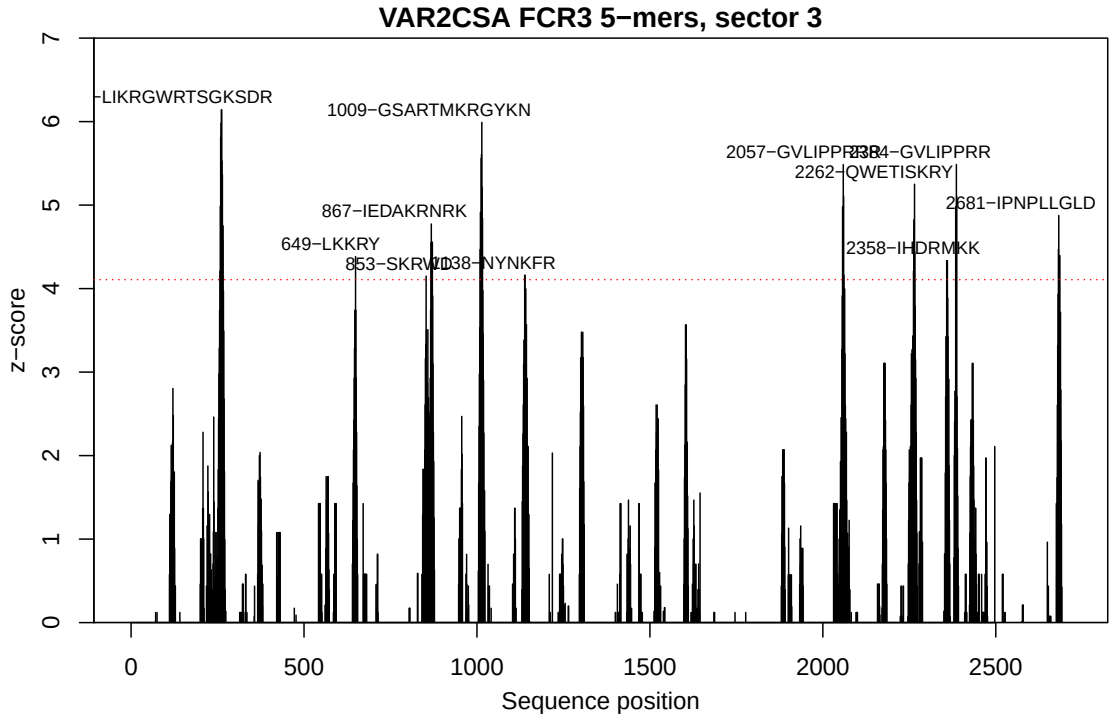
Table 5.3.3: PepChipper-1.0 epitope list for VAR2CSA FCR3

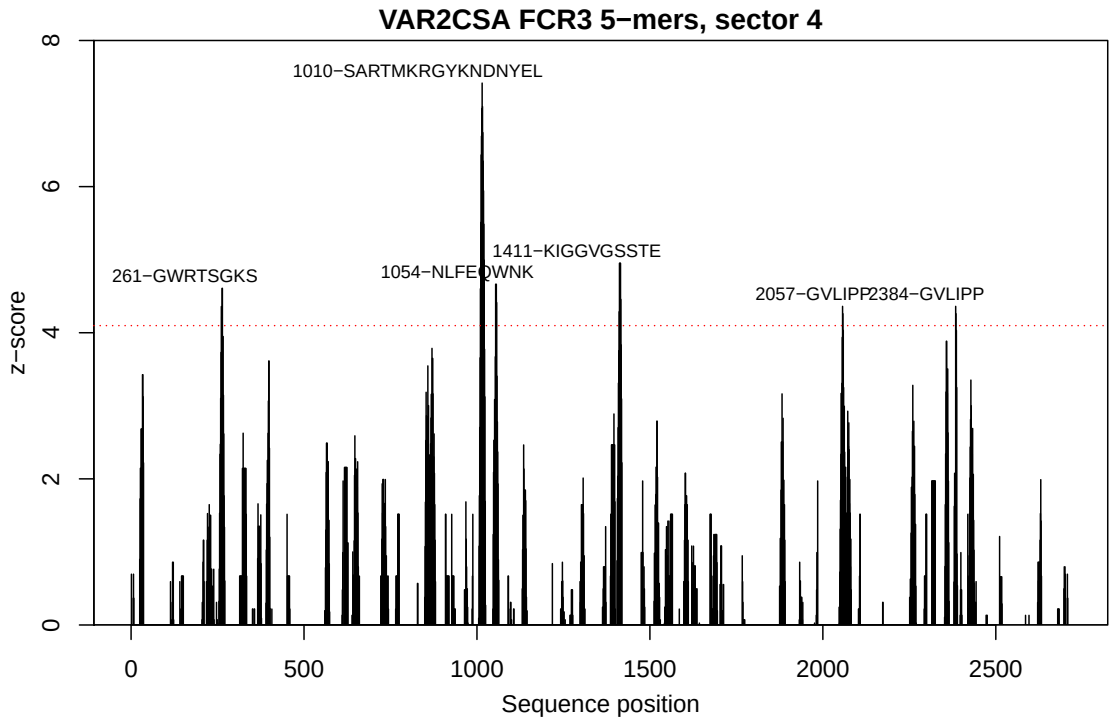
5.3.2 K-mer plots for FCR3 and 3D7

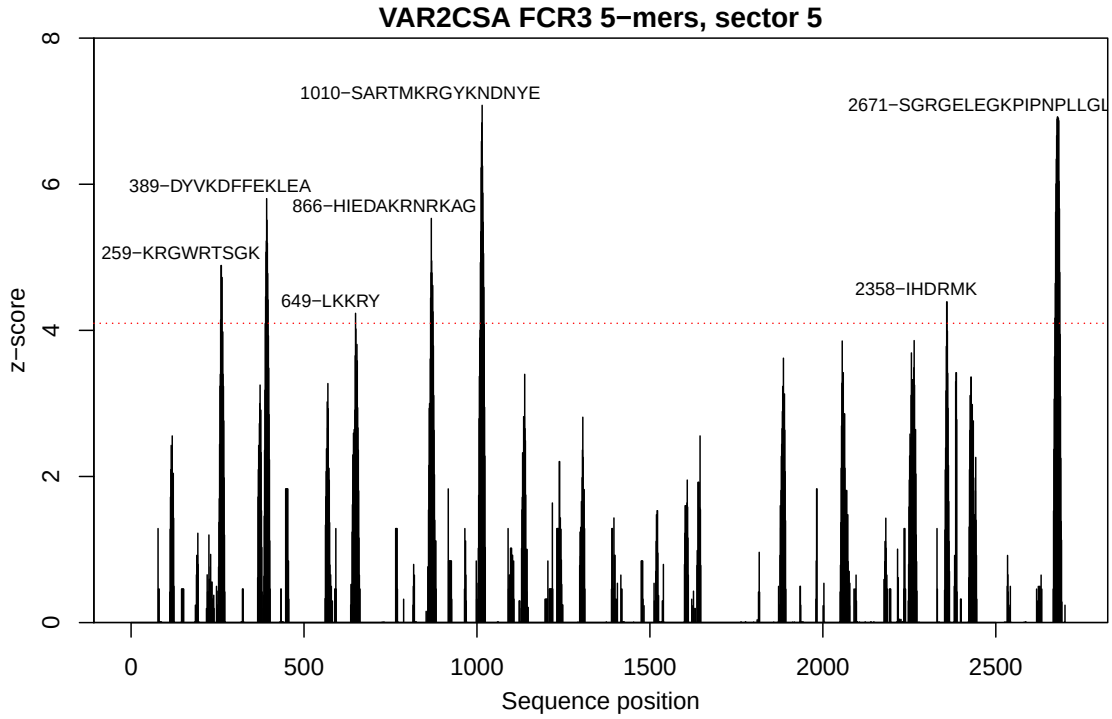
Plots of VAR2CSA FCR3 epitopes identified using the k-mer approach

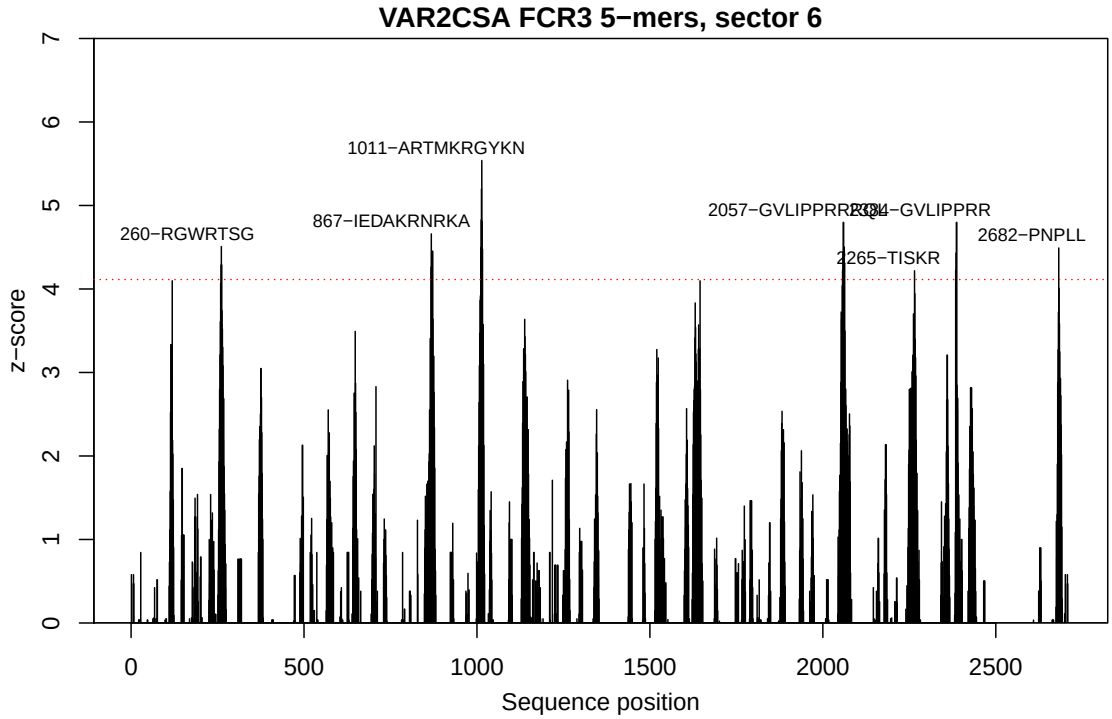

VAR2CSA FCR3 5−mers, sector 1
386−RYDDYVKDFFEKLEA
1008−CGSARTMKRGYKNDNYE
z−score
0
2
4
6
8
0
500
1000
1500
2000
2500
Sequence position

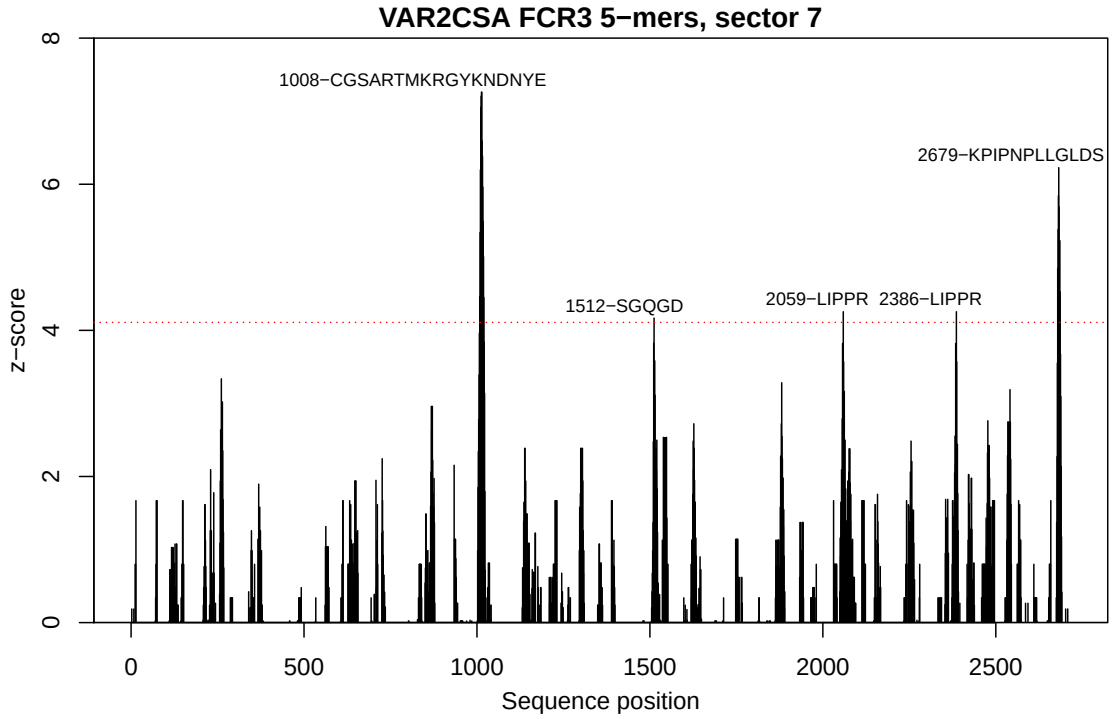

VAR2CSA FCR3 5−mers, sector 2
265−SGKSDRK
260−RGWRTS
545−EGKNLKKRY
868−EDAKRNR
2355−YKNIHDRM
z−score
0
1
2
3
4
5
6
0
500
1000
1500
2000
2500
Sequence position

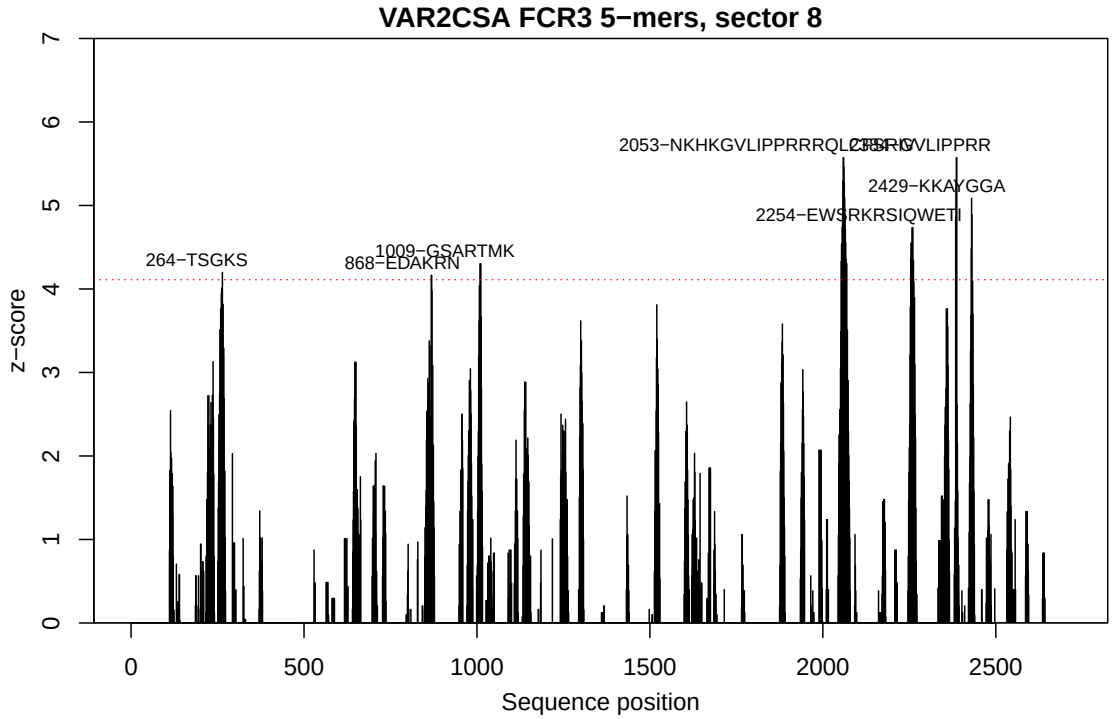

VAR2CSA FCR3 5−mers, sector 3
−LIKRGWRTSGKSDR
1009−GSARTMKRGYKN
867−IEDAKRNRK
649−LKKRY
853−SKRV
1038−NYNKFR
2057−GVLIPPR
2384−GVLIPPRR
2262−QWETISKRY
2681−IPNPLLGLD
2358−IHDRMKK
z−score
Sequence position
0
500
1000
1500
2000
2500
1
2
3
4
5
6
7

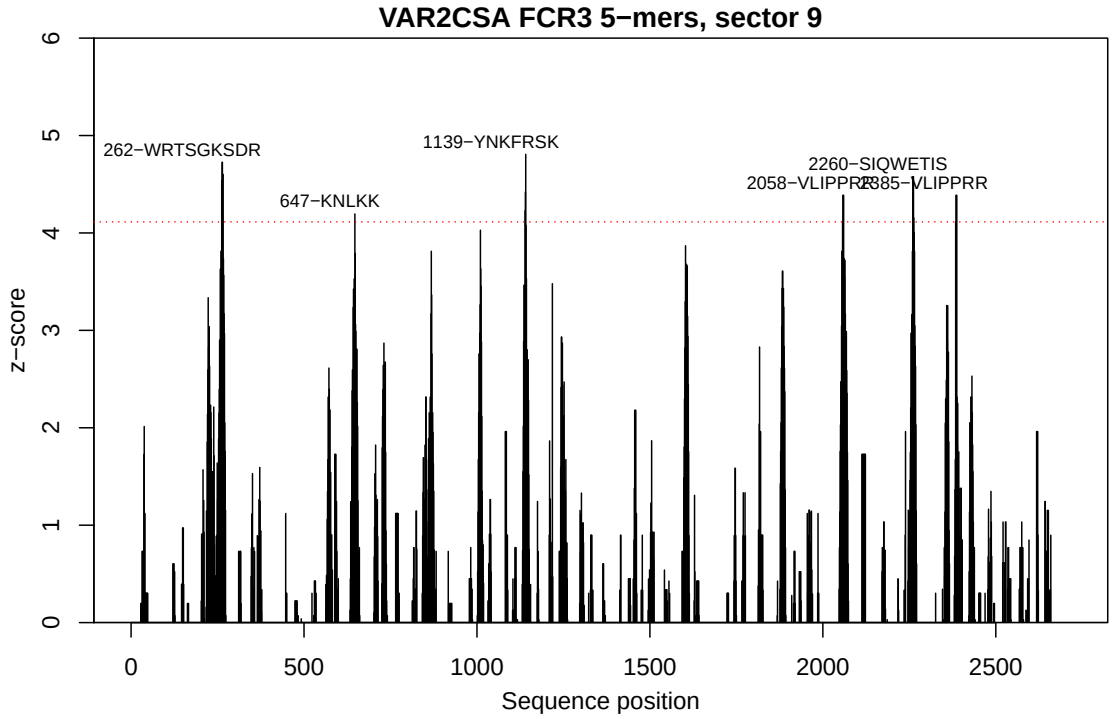

VAR2CSA FCR3 5−mers, sector 4
1010−SARTMKRGYKNDNYEL
261−GWRTSGKS
1411−KIGGVGSSTE
1054−NLFEQWNK
2057−GVLIPP
2384−GVLIPP
z−score
Sequence position
0
500
1000
1500
2000
2500
2
4
6
8

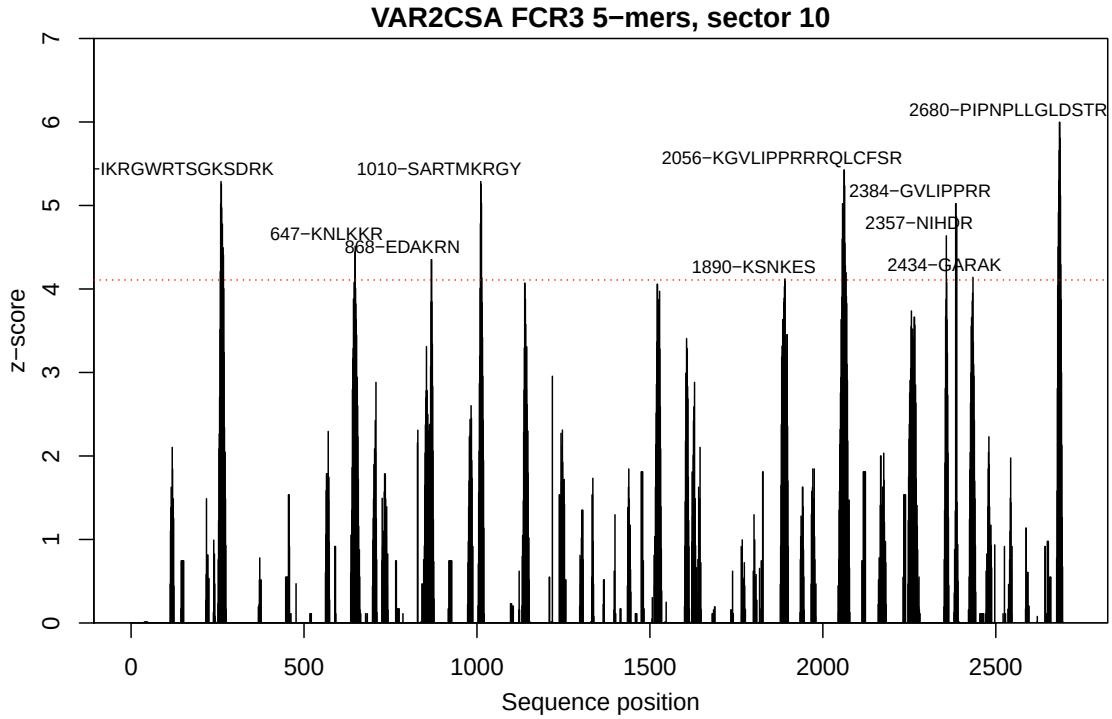

VAR2CSA FCR3 5−mers, sector 5
z−score
Sequence position
1010−SARTMKRGYKNDNYE
2671−SGRGELEGKPIPNPLLGL
389−DYVKDFFEKLEA
866−HIEDAKRNRKAG
259−KRGWRTSGK
649−LKKRY
2358−IHDRMK

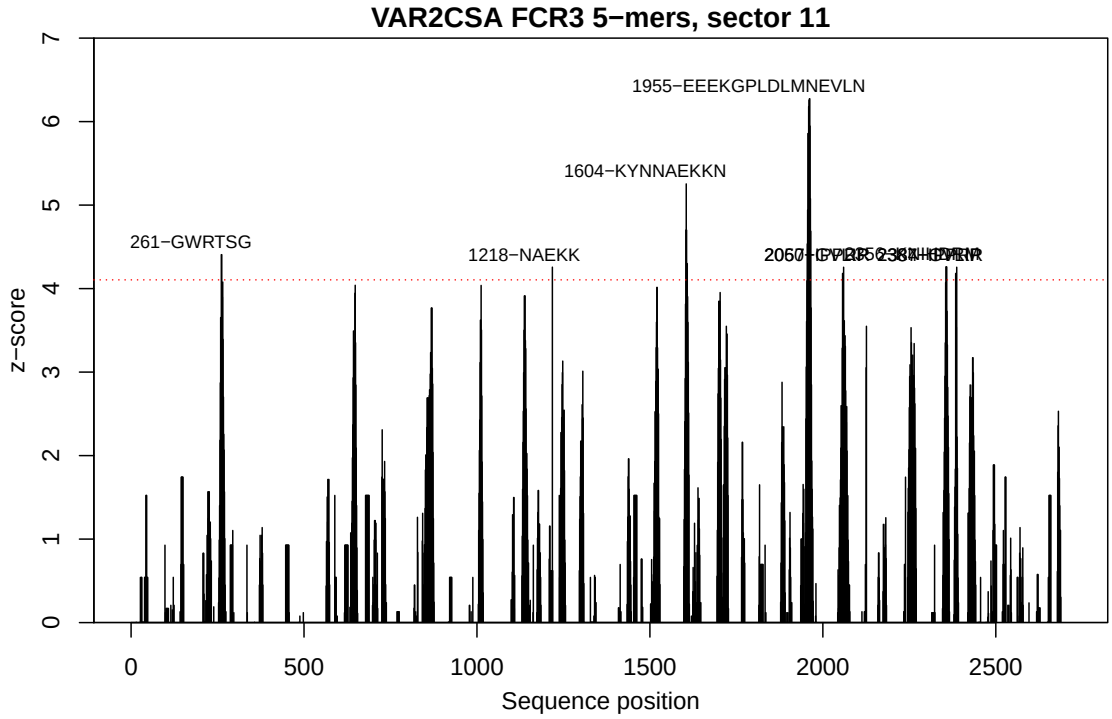

VAR2CSA FCR3 5−mers, sector 6
z−score
Sequence position
1011−ARTMKRGYKN
867−IEDAKRNRKA
260−RGWRTSG
2057−GVLIPPRR
2265−TISKR
2682−PNPLL

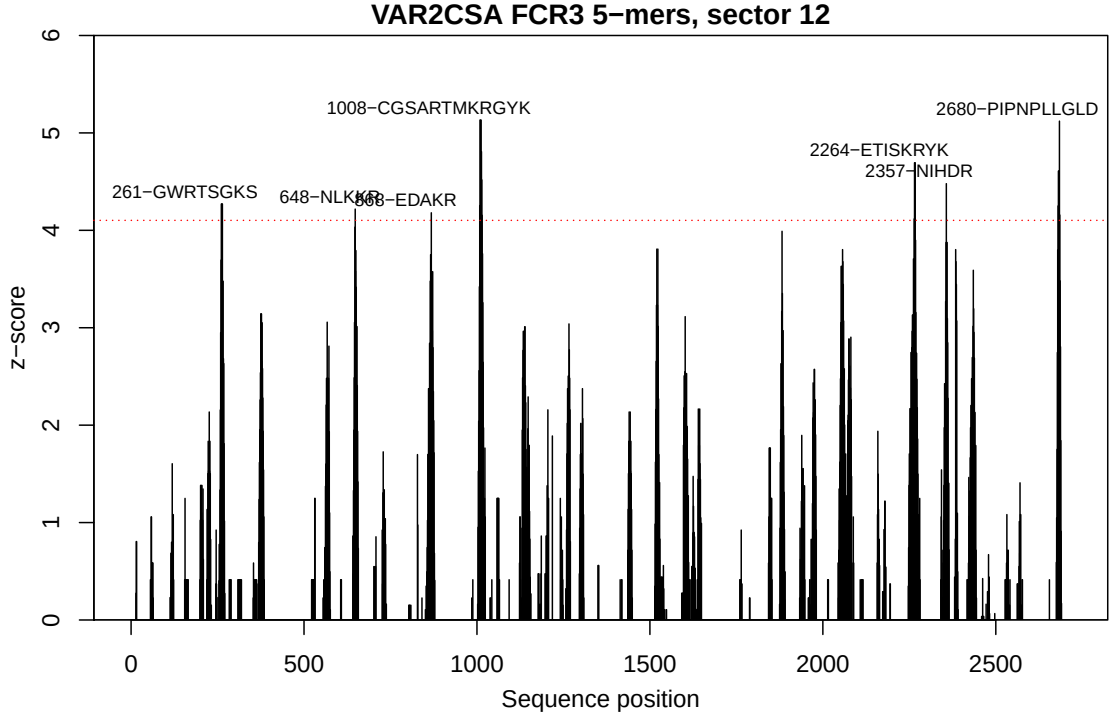

VAR2CSA FCR3 5−mers, sector 7
1008−CGSARTMKRGYKNDNYE
2679−KPIPNPLLGLDS
1512−SGQGD
2059−LIPPR
2386−LIPPR
z−score
Sequence position

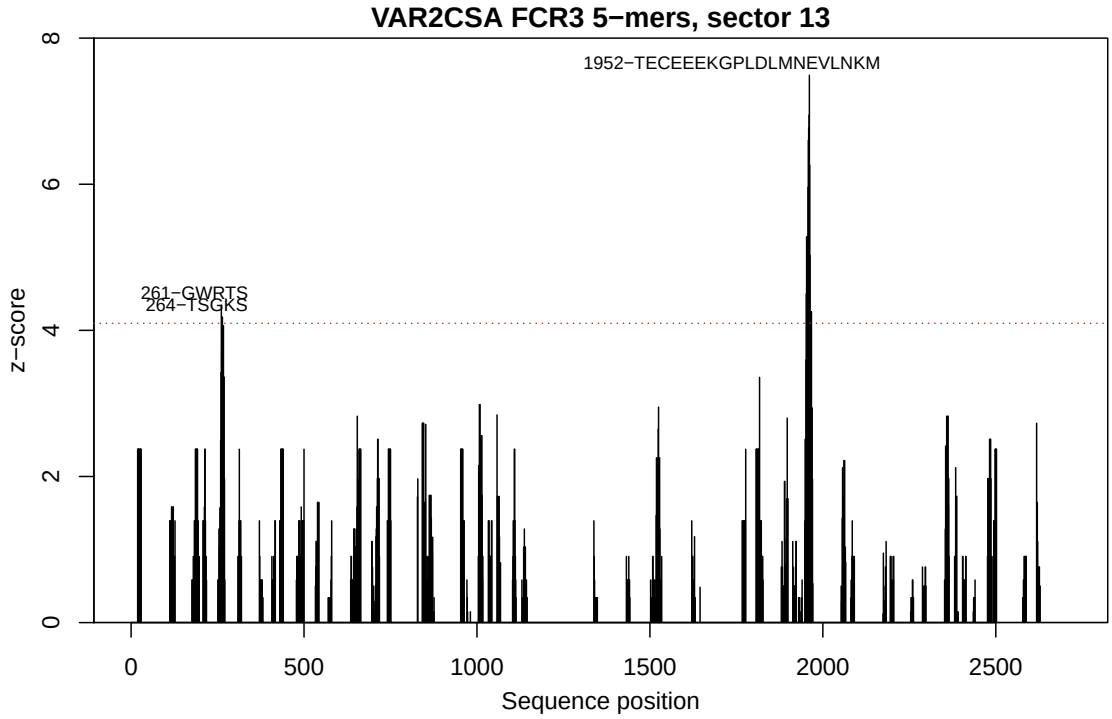

VAR2CSA FCR3 5−mers, sector 8
2053−NKHKGVLIPPRRRQL
2429−KKAYGGA
2254−EWSRKRSIQWETI
264−TSGKS
1009−GSARTMK
868−EDAKRN
z−score
Sequence position

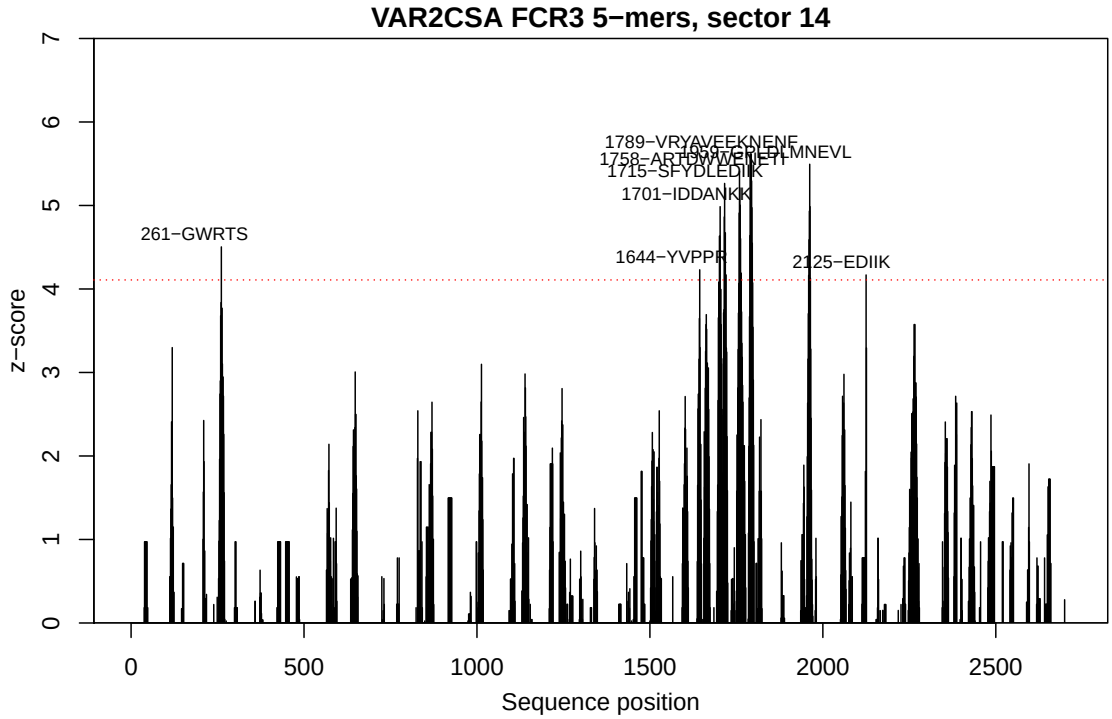

VAR2CSA FCR3 5−mers, sector 9
z−score
Sequence position
262−WRTSGKSDR
647−KNLKK
1139−YNKFRSK
2058−VLIPPRR
2260−SIQWETIS
2385−VLIPPRR
0
1
2
3
4
5
6
0
500
1000
1500
2000
2500

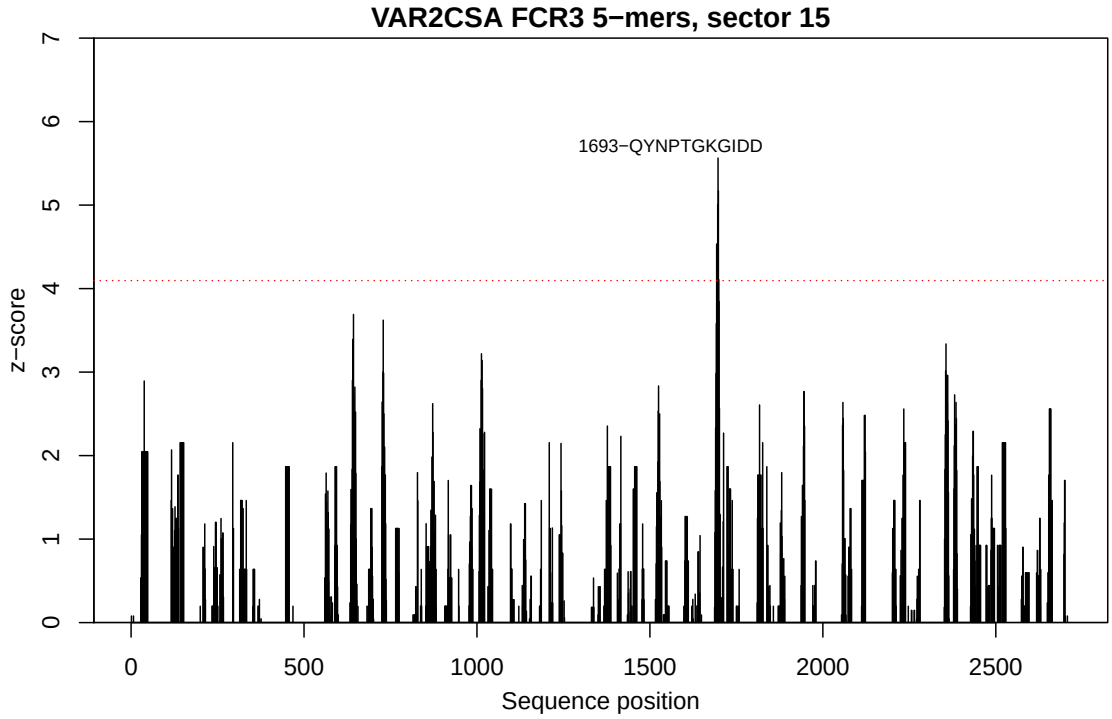

VAR2CSA FCR3 5−mers, sector 10
z−score
Sequence position
−IKRGWRTSGKSDRK
647−KNLKKR
868−EDAKRN
1010−SARTMKRGY
1890−KSNKES
2056−KGVLIPPRRRQLCFSR
2384−GVLIPPRR
2357−NIHDR
2434−GARAK
2680−PIPNPLLGLDSTR
0
1
2
3
4
5
6
7
0
500
1000
1500
2000
2500

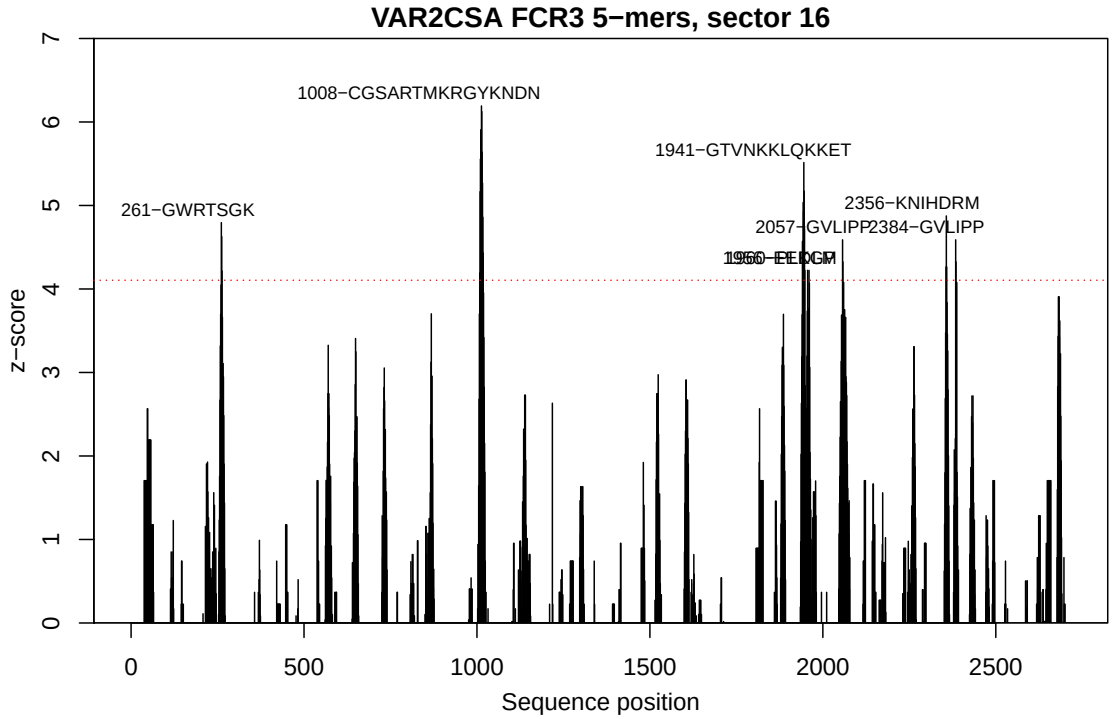

VAR2CSA FCR3 5−mers, sector 11
z−score
Sequence position
261−GWRTSG
1218−NAEKK
1604−KYNNAEKKN
1955−EEEKGPLDLMNEVLN
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6
7

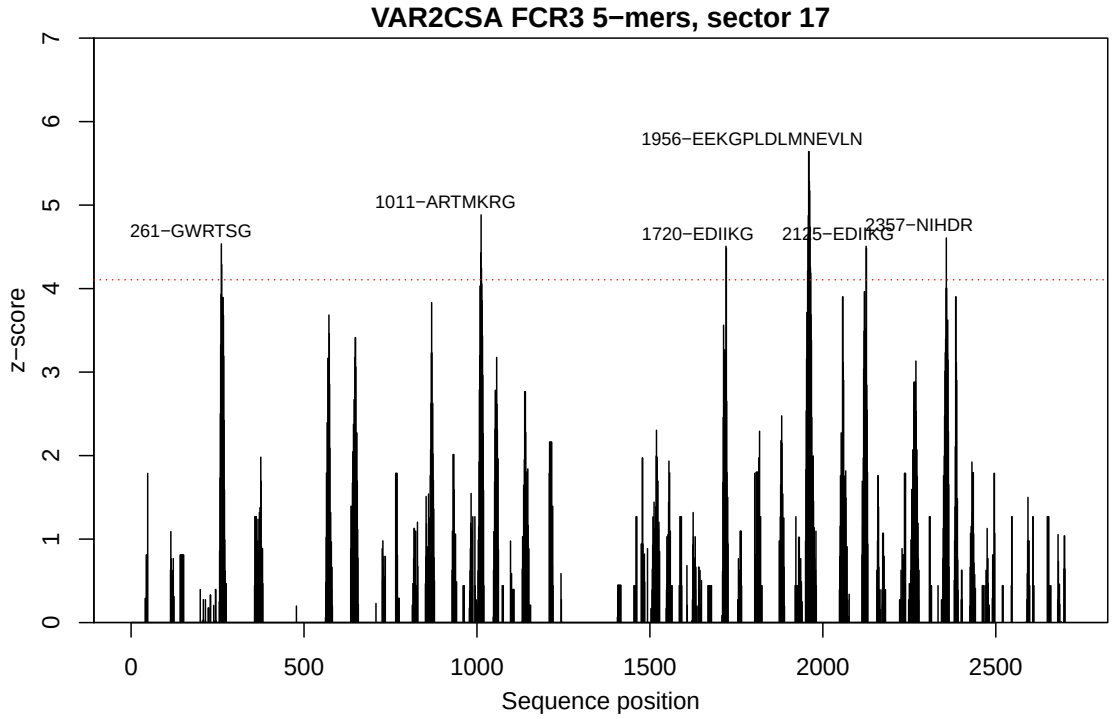

VAR2CSA FCR3 5−mers, sector 12
z−score
Sequence position
261−GWRTSGKS
648−NLKKR
868−EDAKR
1008−CGSARTMKRGYK
2264−ETISKRYK
2357−NIHDR
2680−PIPNPLLGLD
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6

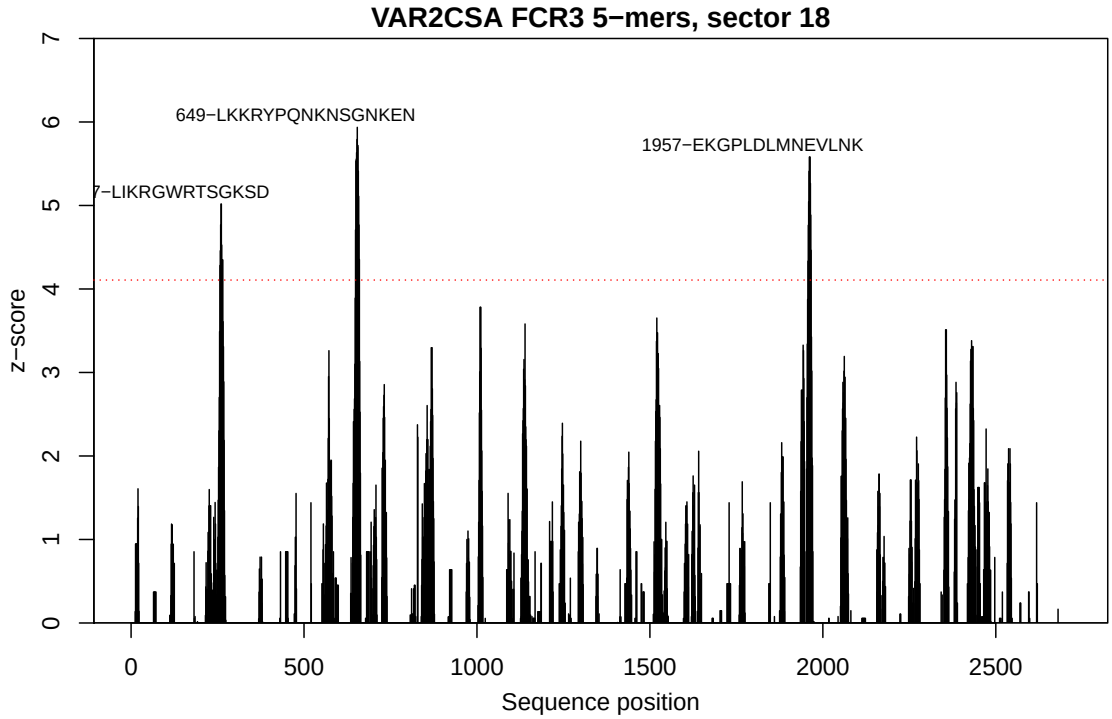

VAR2CSA FCR3 5−mers, sector 13
z−score
Sequence position
1952−TECEEEKGPLDLMNEVLNKM
261−GWRTS
264−TSGKS
0
2
4
6
8
0
500
1000
1500
2000
2500

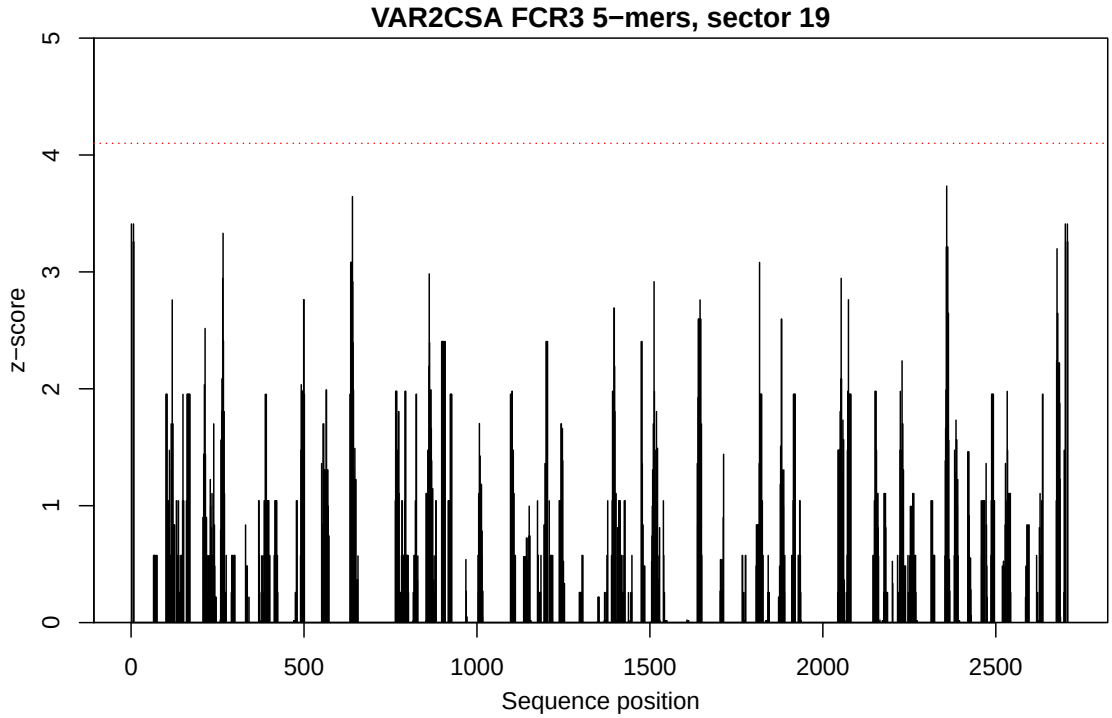

VAR2CSA FCR3 5−mers, sector 14
z−score
Sequence position
1789−VRYAVEEKNENE
1758−ARTDWVEMETI
1715−SFYDLEDIIK
1701−IDDANKK
1644−YVPPR
261−GWRTS
2125−EDIIK
0
1
2
3
4
5
6
7
0
500
1000
1500
2000
2500

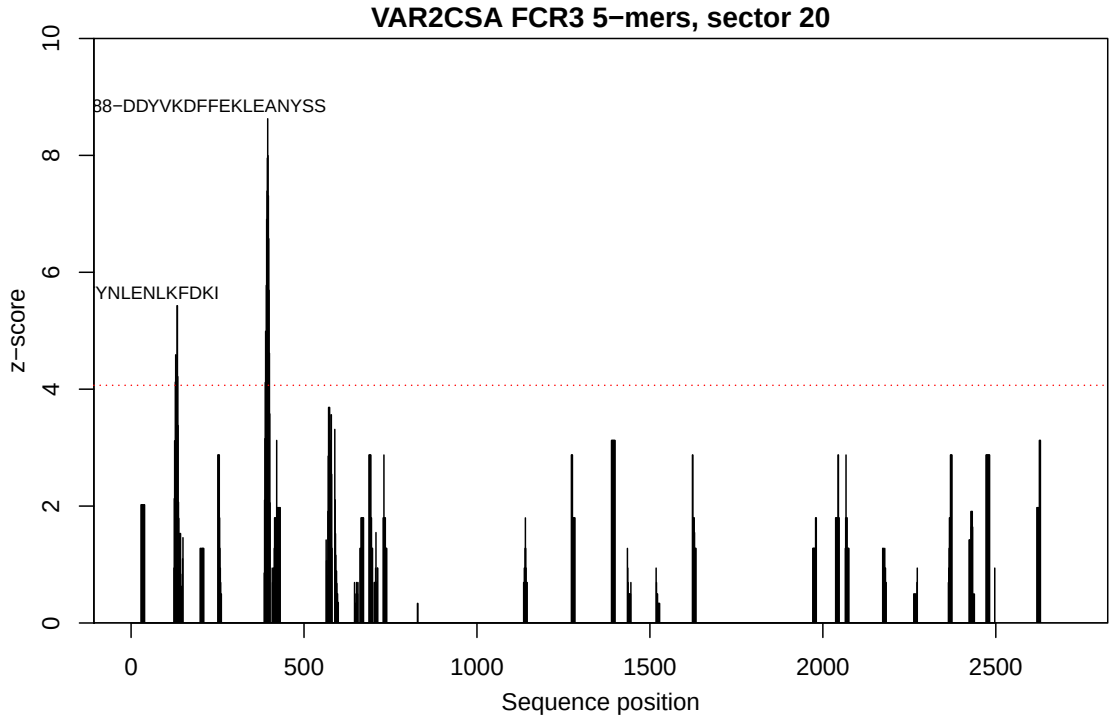

VAR2CSA FCR3 5−mers, sector 15
1693−QYNPTGKGIDD
z−score
Sequence position
0
500
1000
1500
2000
2500
1
2
3
4
5
6
7

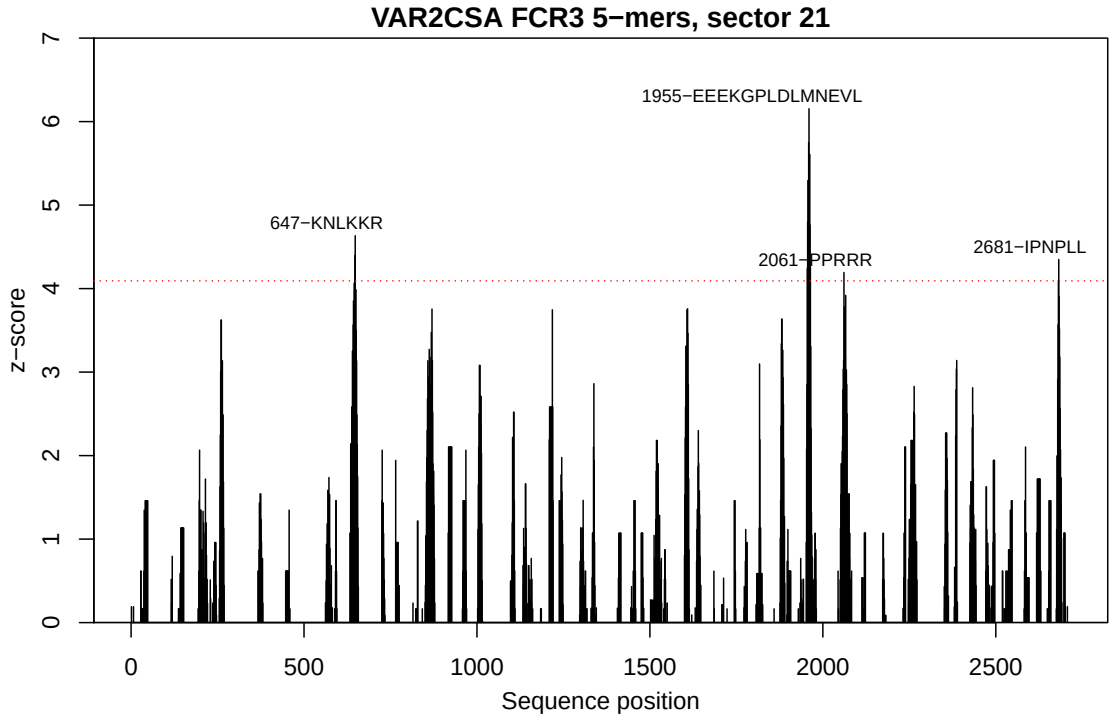

VAR2CSA FCR3 5−mers, sector 16
1008−CGSARTMKRGYKNDN
1941−GTVNKKLQKKET
261−GWRTSGK
2356−KNIHDRM
2057−GVLIPP
2384−GVLIPP
z−score
Sequence position
0
500
1000
1500
2000
2500
1
2
3
4
5
6
7

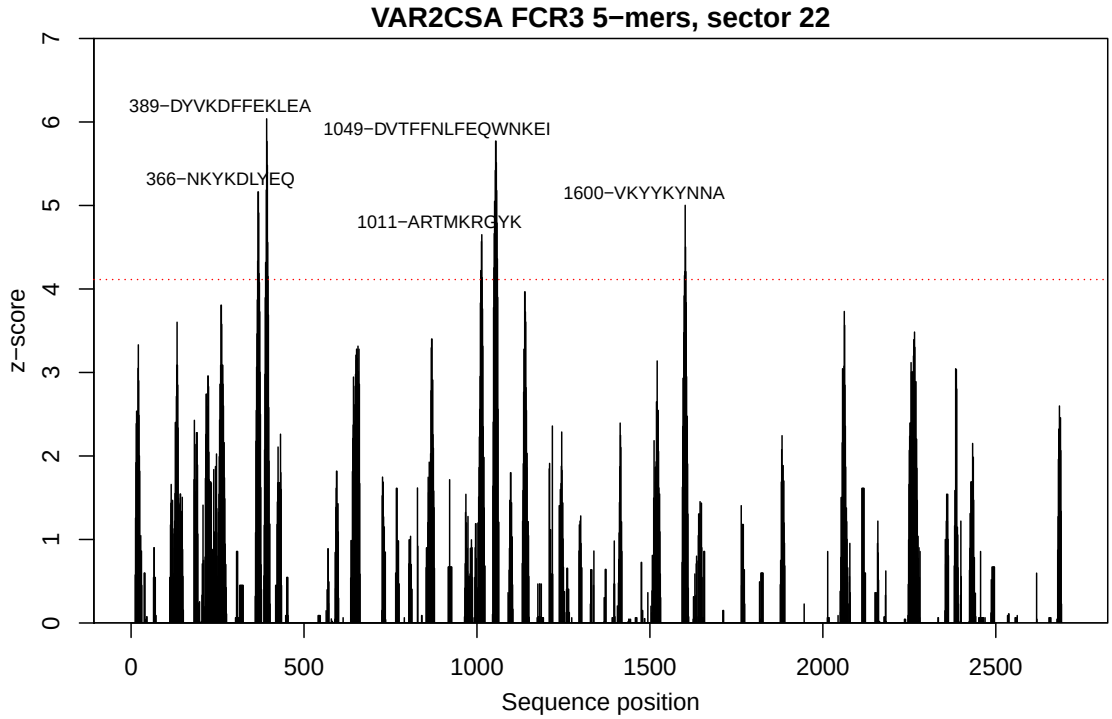

VAR2CSA FCR3 5−mers, sector 17
z−score
Sequence position
261−GWRTSG
1011−ARTMKRG
1720−EDIIKG
1956−EEKGPLDLMNEVLN
2125−EDIIKG
2357−NIHDR
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6
7

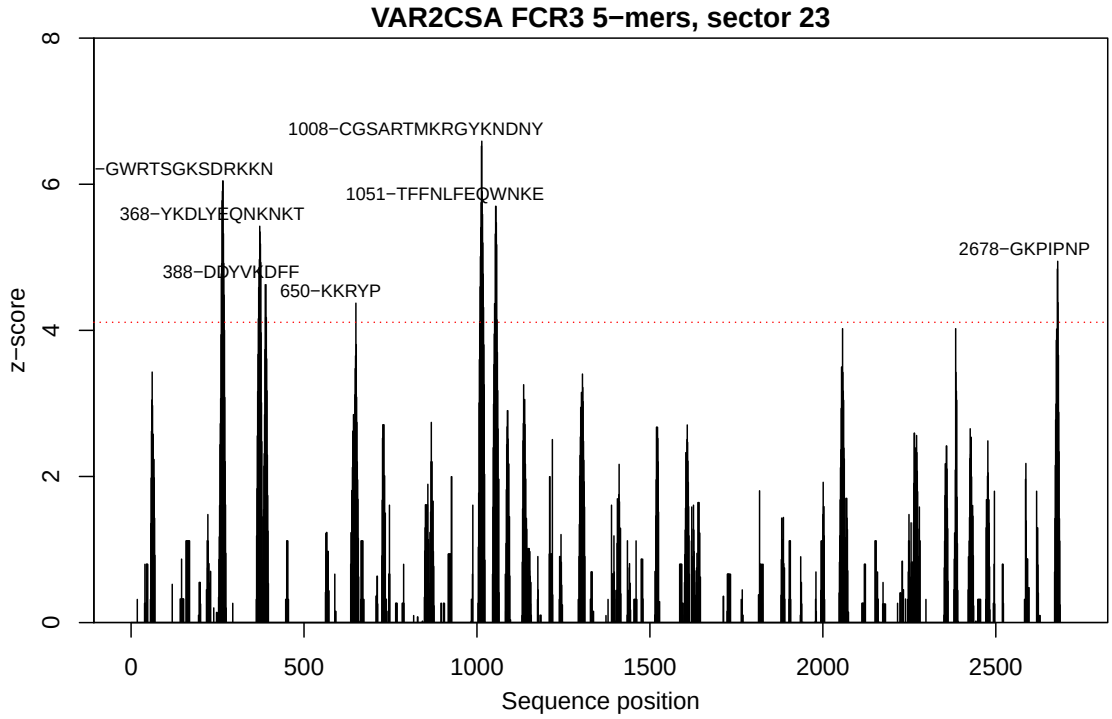

VAR2CSA FCR3 5−mers, sector 18
z−score
Sequence position
649−LKKRYPQNKNSGNKEN
1957−EKGPLDLMNEVLNK
7−LIKRGWRTSGKSD
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6
7

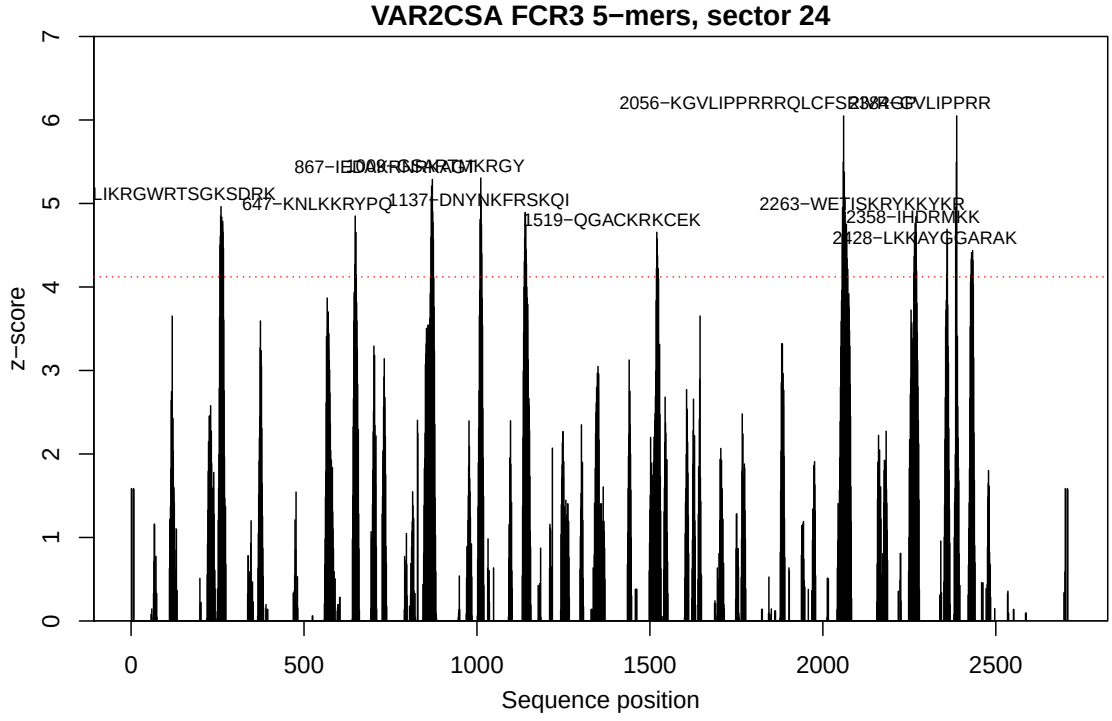

VAR2CSA FCR3 5−mers, sector 19
z−score
0
1
2
3
4
5
0
500
1000
1500
2000
2500
Sequence position


VAR2CSA FCR3 5−mers, sector 20
88−DDYVKDFFEKLEANYSS
YNLENLKFDKI
z−score
0
2
4
6
8
10
0
500
1000
1500
2000
2500
Sequence position

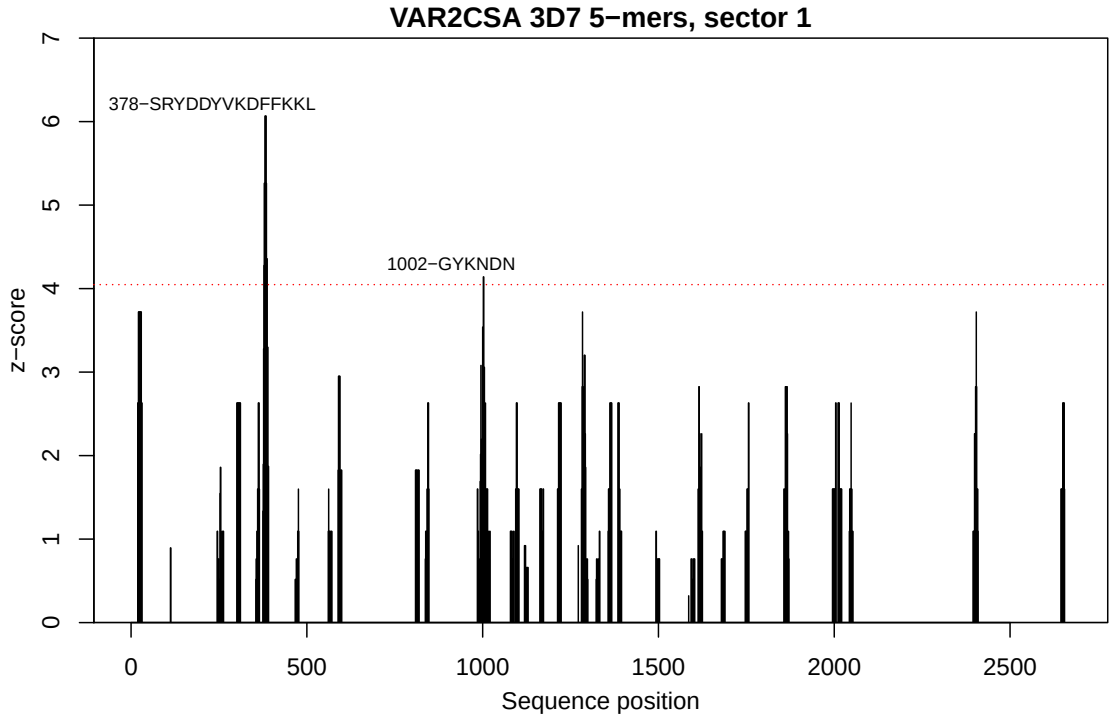

VAR2CSA FCR3 5−mers, sector 21
1955−EEEKGPLDLMNEVL
647−KNLKKR
2061−PPRRR
2681−IPNPLL
z−score
Sequence position
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6
7

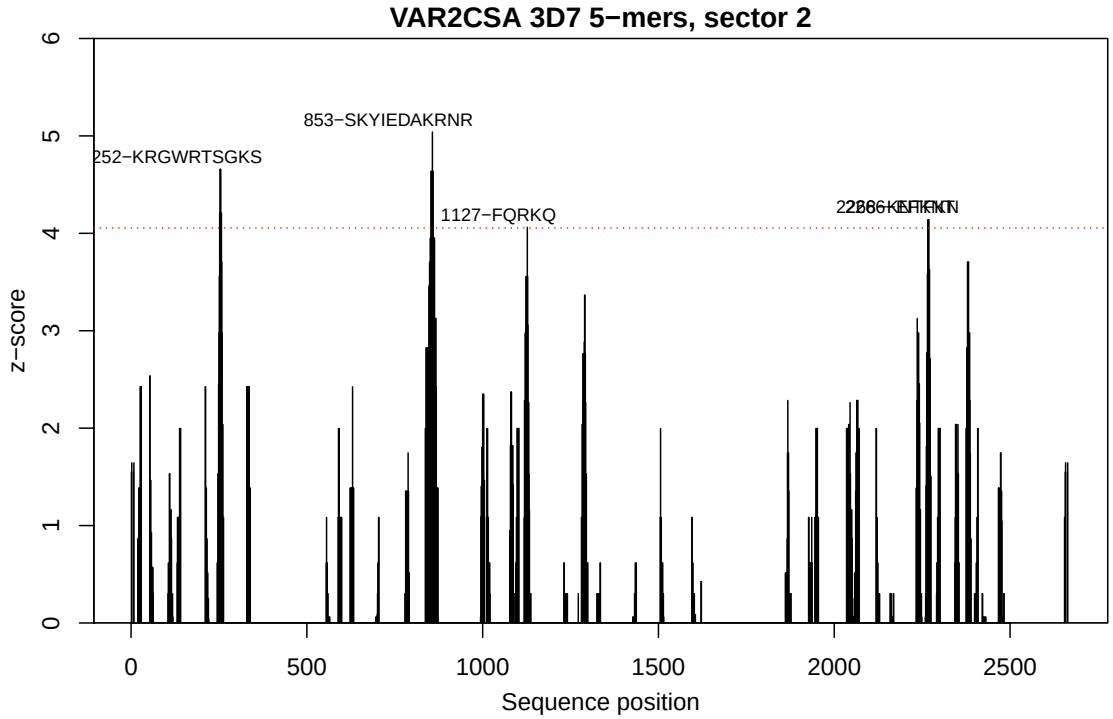

VAR2CSA FCR3 5−mers, sector 22
389−DYVKDFFEKLEA
1049−DVTFFNLFEQWNKEI
366−NKYKDLYEQ
1600−VKYYKYNNA
1011−ARTMKRGYK
z−score
Sequence position
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6
7

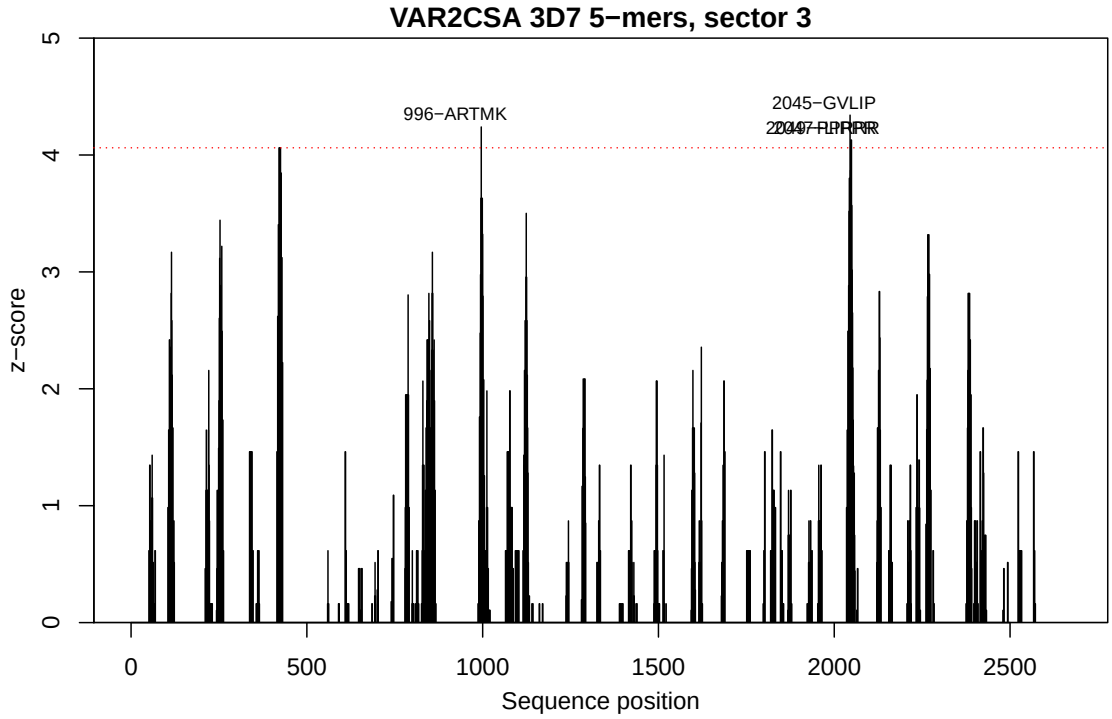

VAR2CSA FCR3 5−mers, sector 23
z−score
Sequence position
1008−CGSARTMKRGYKNDNY
−GWRTSGKSDRKKN
1051−TFFNLFEQWNKE
368−YKDLYEQNKNKT
2678−GKPIPNP
388−DDYVKDFF
650−KKRYP
0
2
4
6
8
0
500
1000
1500
2000
2500

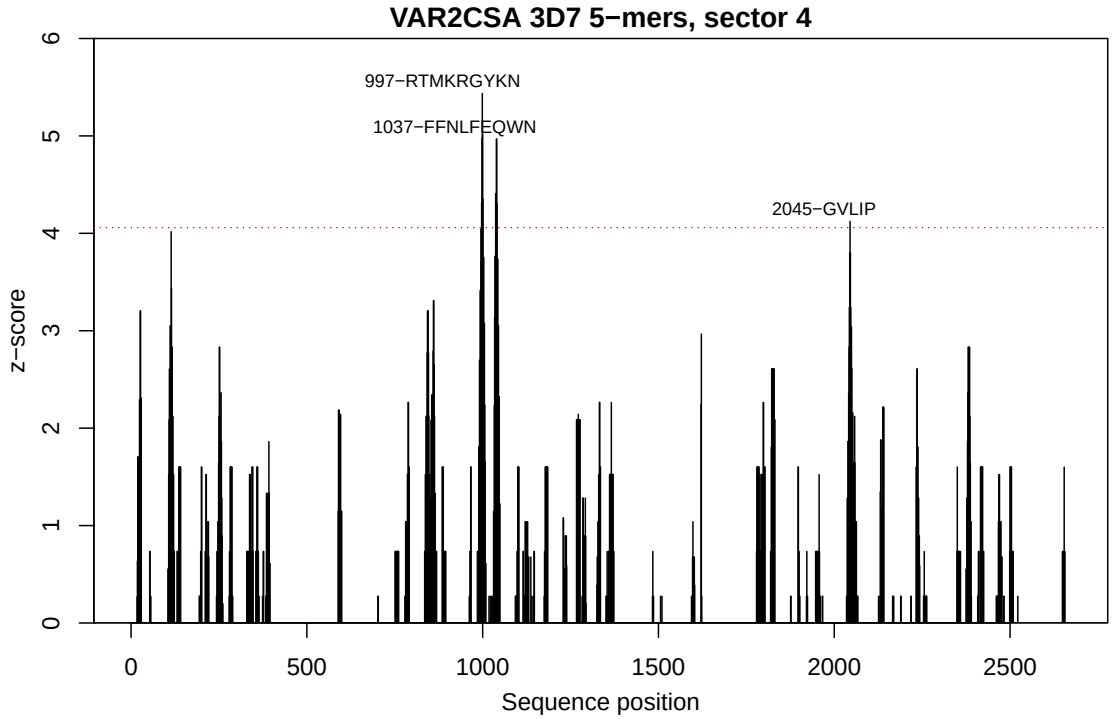

VAR2CSA FCR3 5−mers, sector 24
z−score
Sequence position
2056−KGVLIPPRRRQLCFS
LIKRGWRTSGKSDRK
647−KNLKKRYPQ
1137−DNYNKFRSKQI
1519−QGACKRKCEK
2263−WETISKRYKKYKR
2358−IHDRMKK
2428−LKKAYGGARAK
0
1
2
3
4
5
6
7
0
500
1000
1500
2000
2500

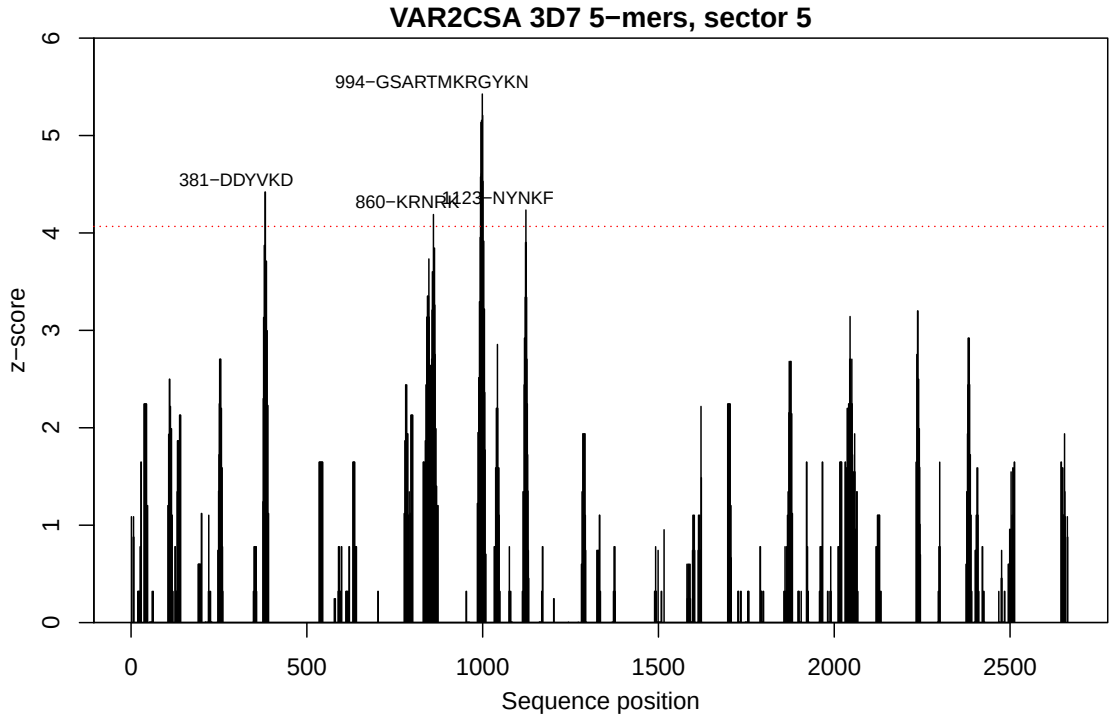
Plots of VAR2CSA 3D7 epitopes identified using the k-mer approach

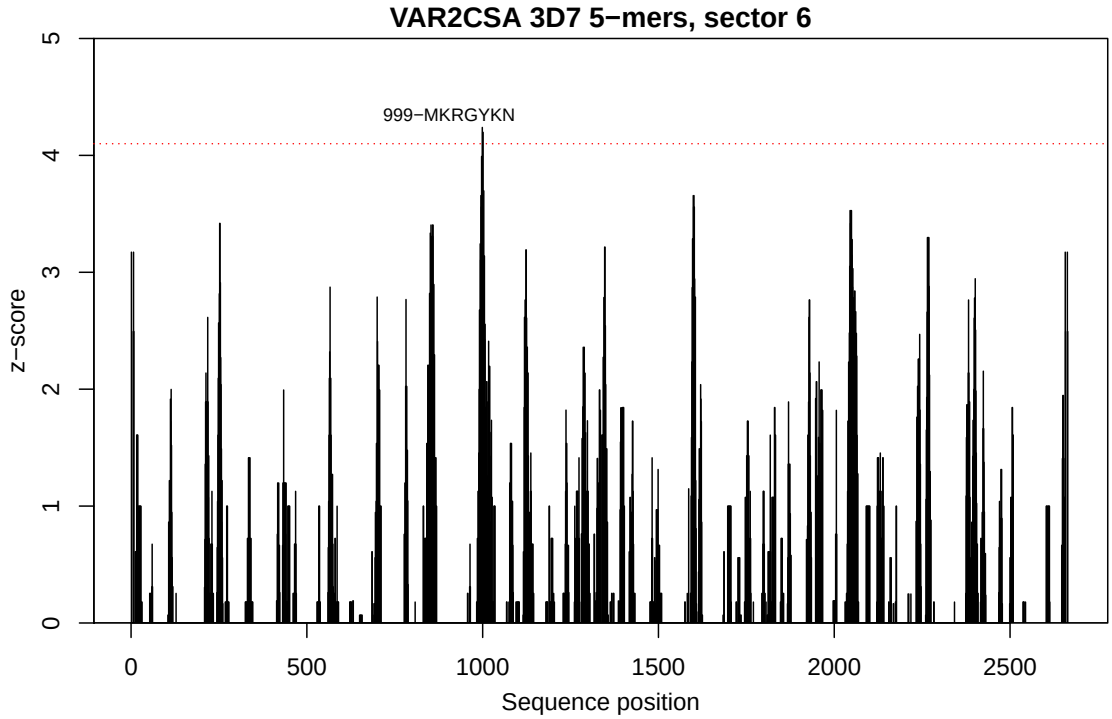

VAR2CSA 3D7 5−mers, sector 1
z−score
Sequence position
378−SRYDDYVKDFFKKL
1002−GYKNDN
0
1
2
3
4
5
6
7
0
500
1000
1500
2000
2500

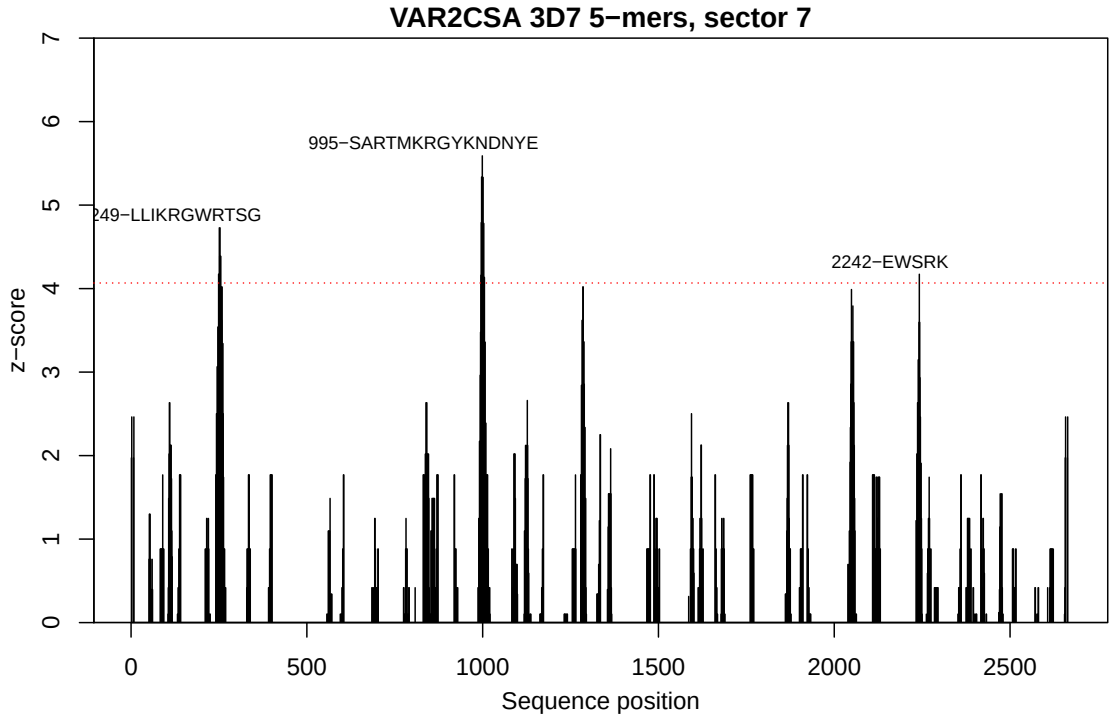

VAR2CSA 3D7 5−mers, sector 2
z−score
Sequence position
853−SKYIEDAKRNR
252−KRGWRTSGKS
1127−FQRKQ
0
1
2
3
4
5
6
0
500
1000
1500
2000
2500

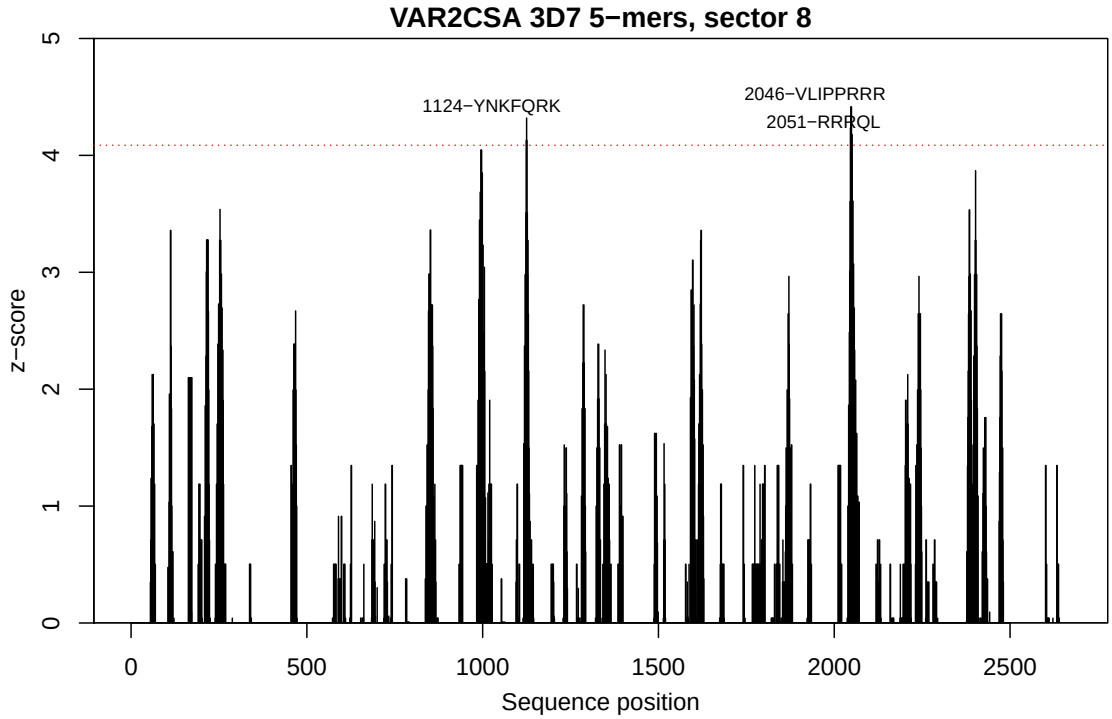

VAR2CSA 3D7 5−mers, sector 3
996−ARTMK
2045−GVLIP
z−score
Sequence position
0
500
1000
1500
2000
2500
0
1
2
3
4
5

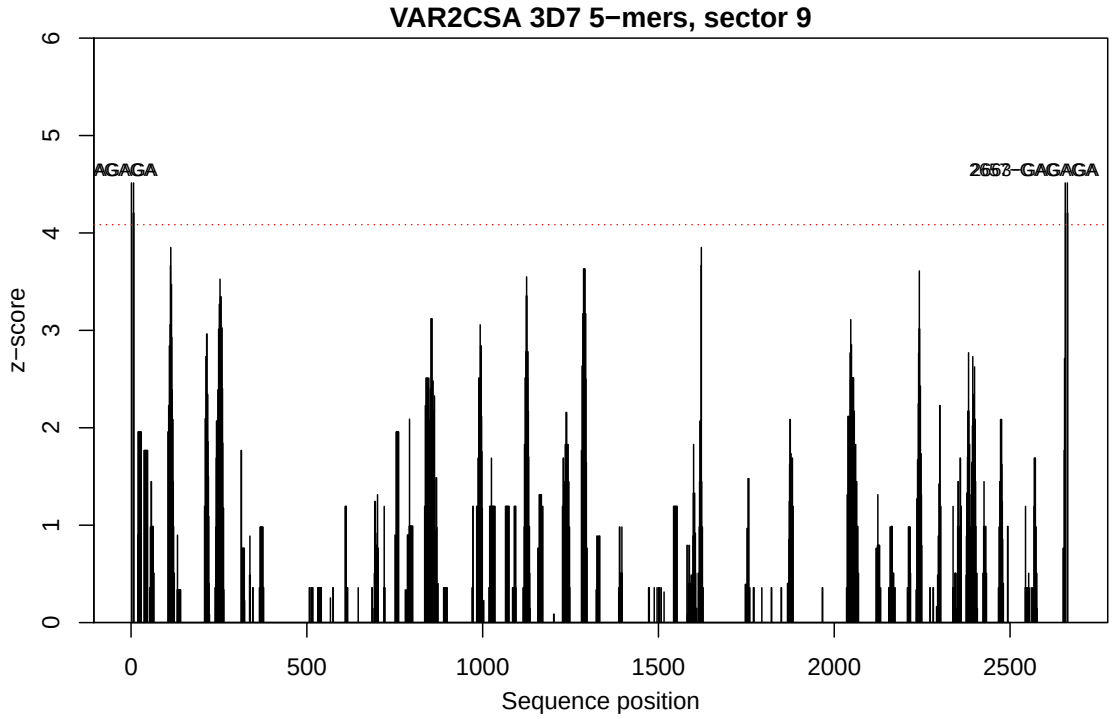

VAR2CSA 3D7 5−mers, sector 4
997−RTMKRGYKN
1037−FFNLFEQWN
2045−GVLIP
z−score
Sequence position
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6

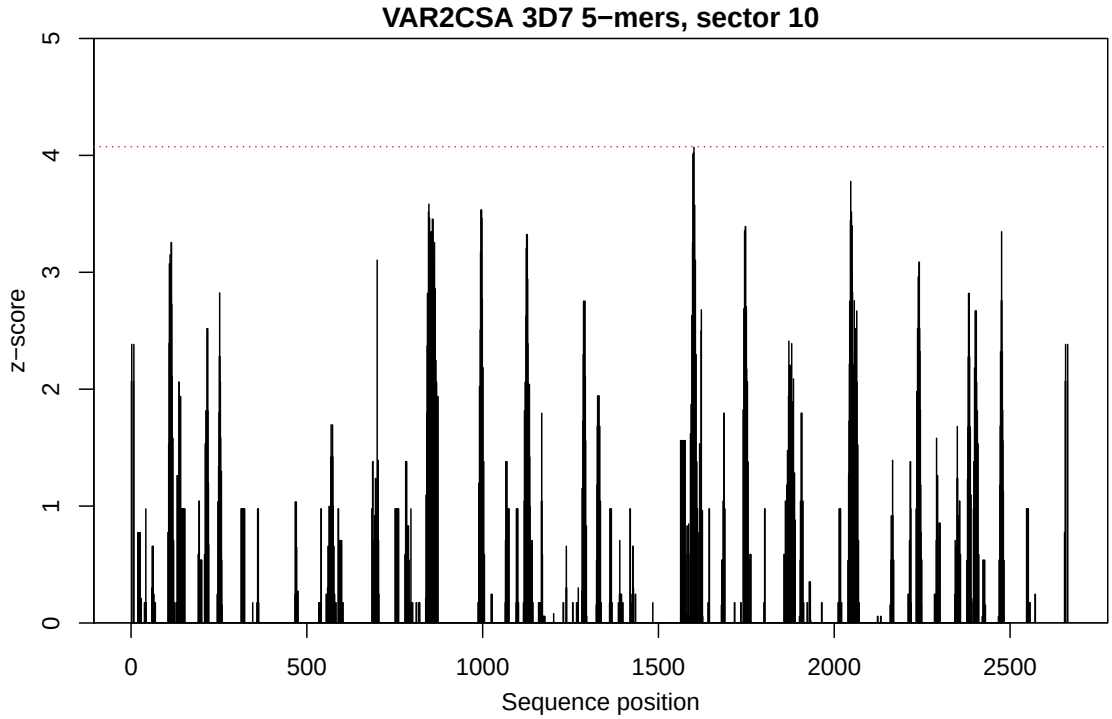

VAR2CSA 3D7 5−mers, sector 5
994−GSARTMKRGYKN
381−DDYVKD
860−KRNRK
1123−NYNKF
z−score
Sequence position
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6

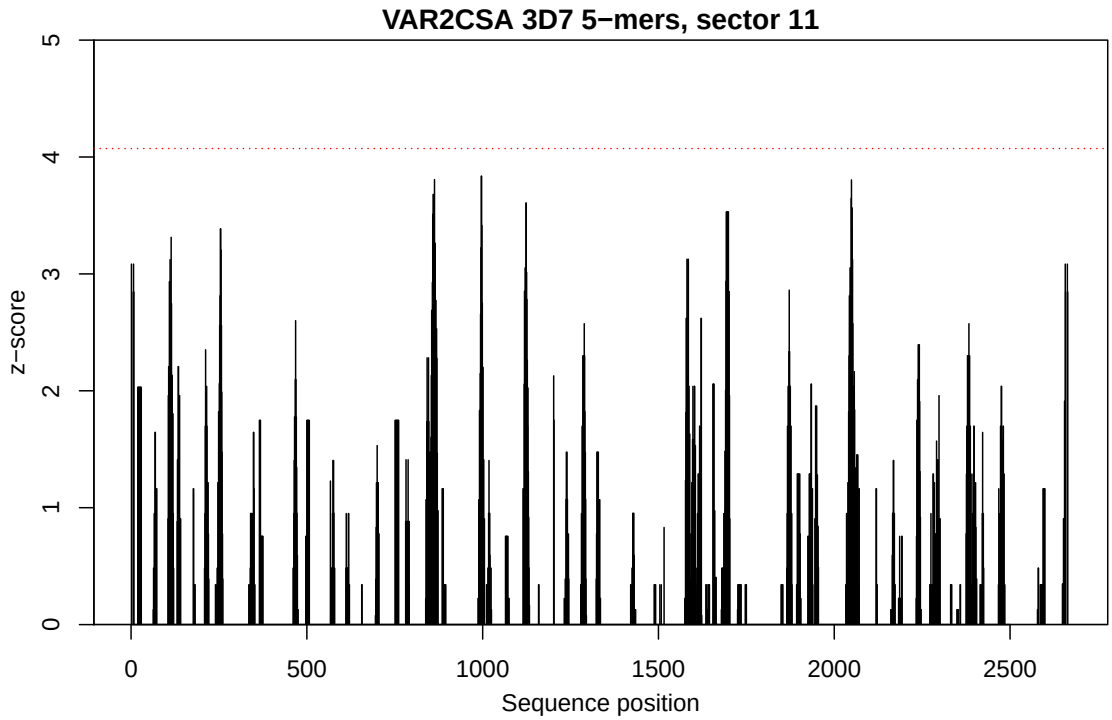

VAR2CSA 3D7 5−mers, sector 6
999−MKRGYKN
z−score
Sequence position
0
500
1000
1500
2000
2500
0
1
2
3
4
5

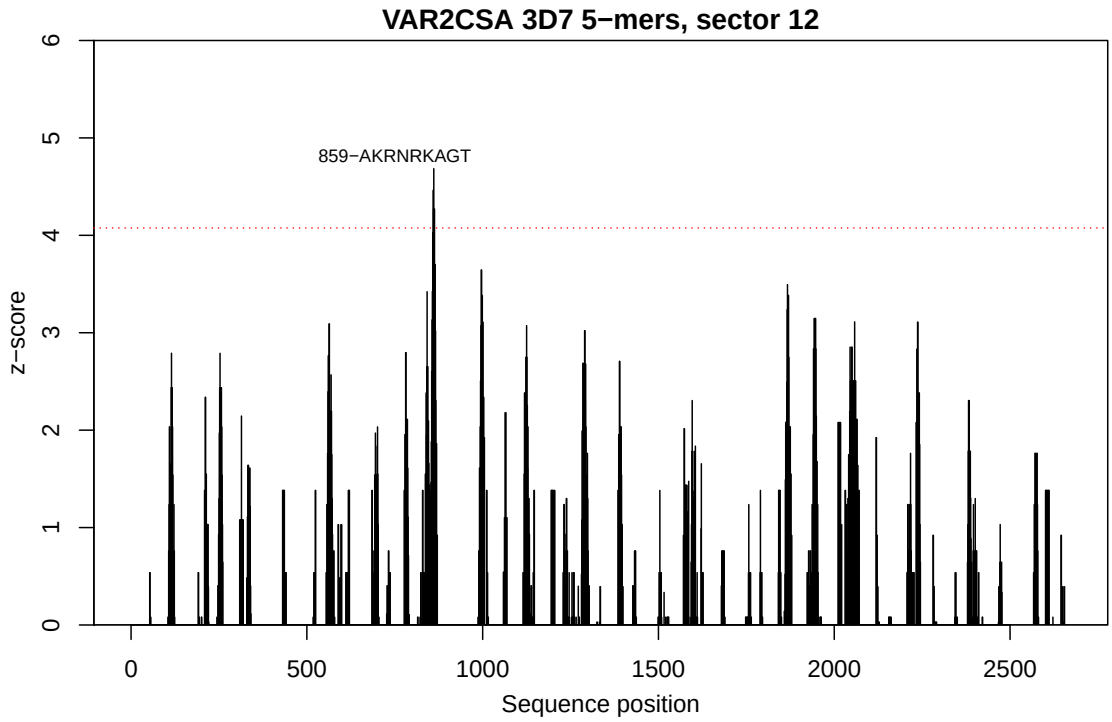

VAR2CSA 3D7 5−mers, sector 7
z−score
Sequence position
995−SARTMKRGYKNDNYE
249−LLIKRGWRTSG
2242−EWSRK
0
1
2
3
4
5
6
7
0
500
1000
1500
2000
2500

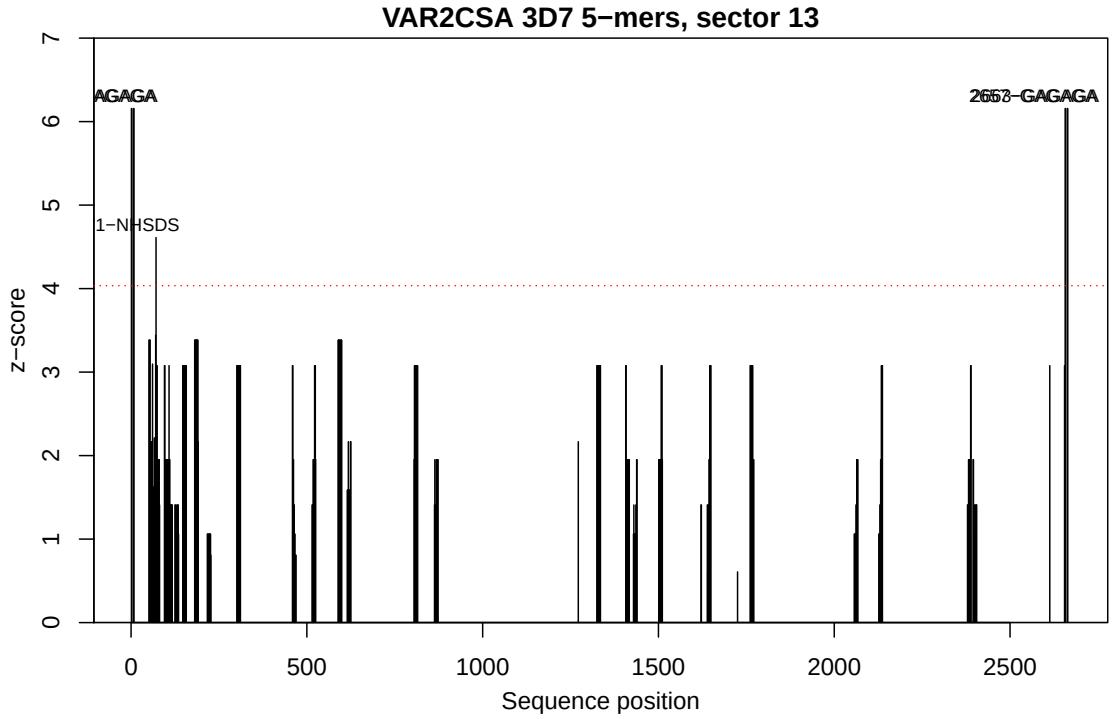

VAR2CSA 3D7 5−mers, sector 8
z−score
Sequence position
1124−YNKFQRK
2046−VLIPPRRR
2051−RRRQL
0
1
2
3
4
5
0
500
1000
1500
2000
2500

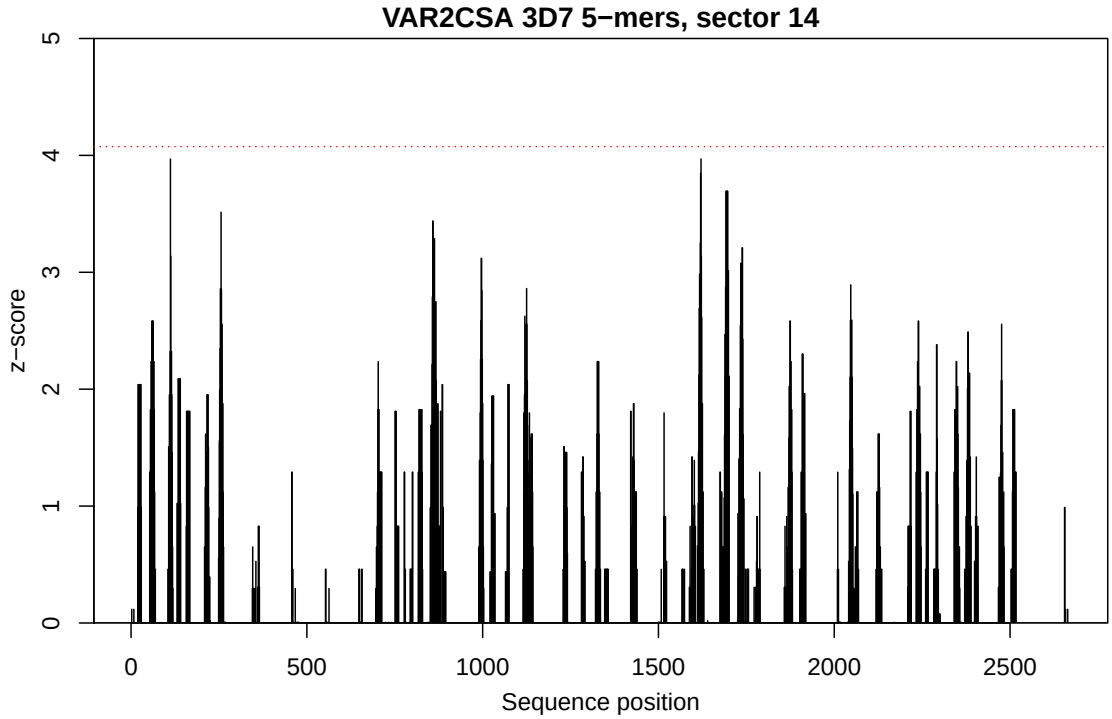

VAR2CSA 3D7 5−mers, sector 9
AGAGA
2663−GAGAGA
z−score
Sequence position
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6

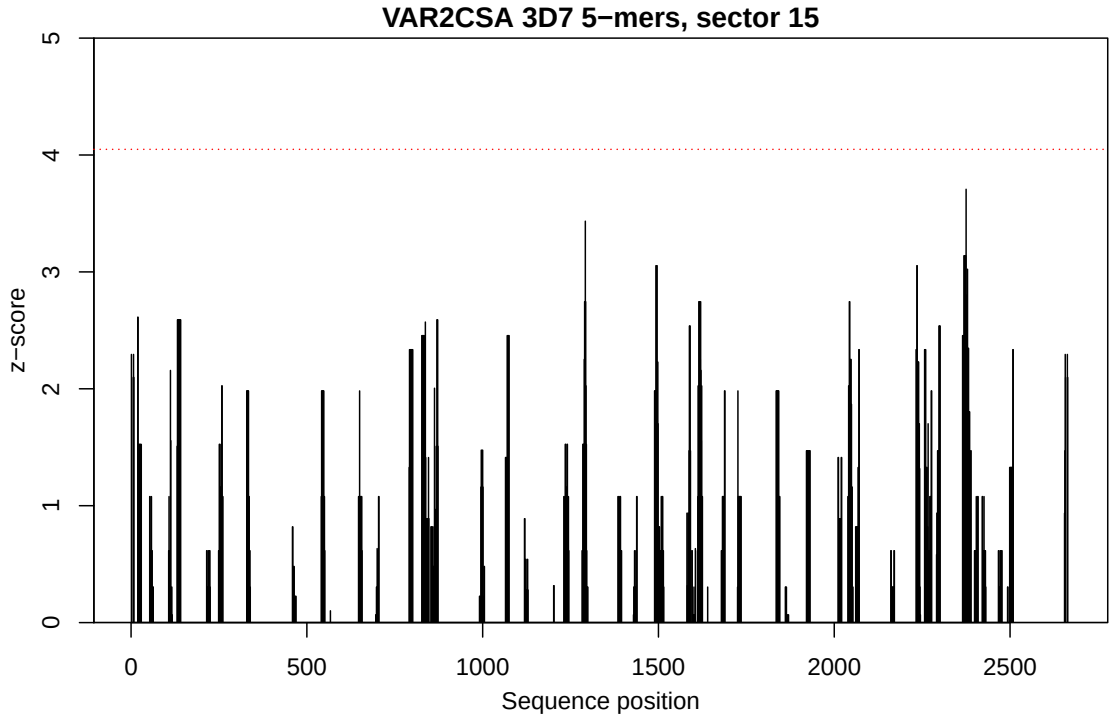

VAR2CSA 3D7 5−mers, sector 10
z−score
Sequence position
0
500
1000
1500
2000
2500
0
1
2
3
4
5

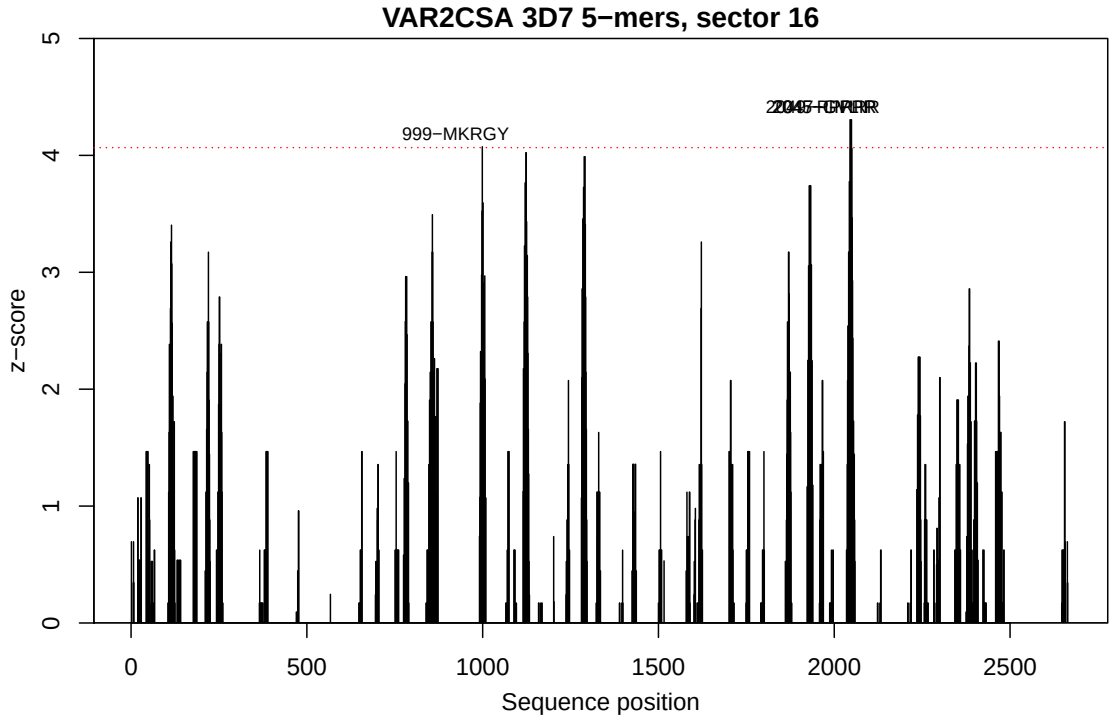

VAR2CSA 3D7 5−mers, sector 11
z−score
0
1
2
3
4
5
0
500
1000
1500
2000
2500
Sequence position

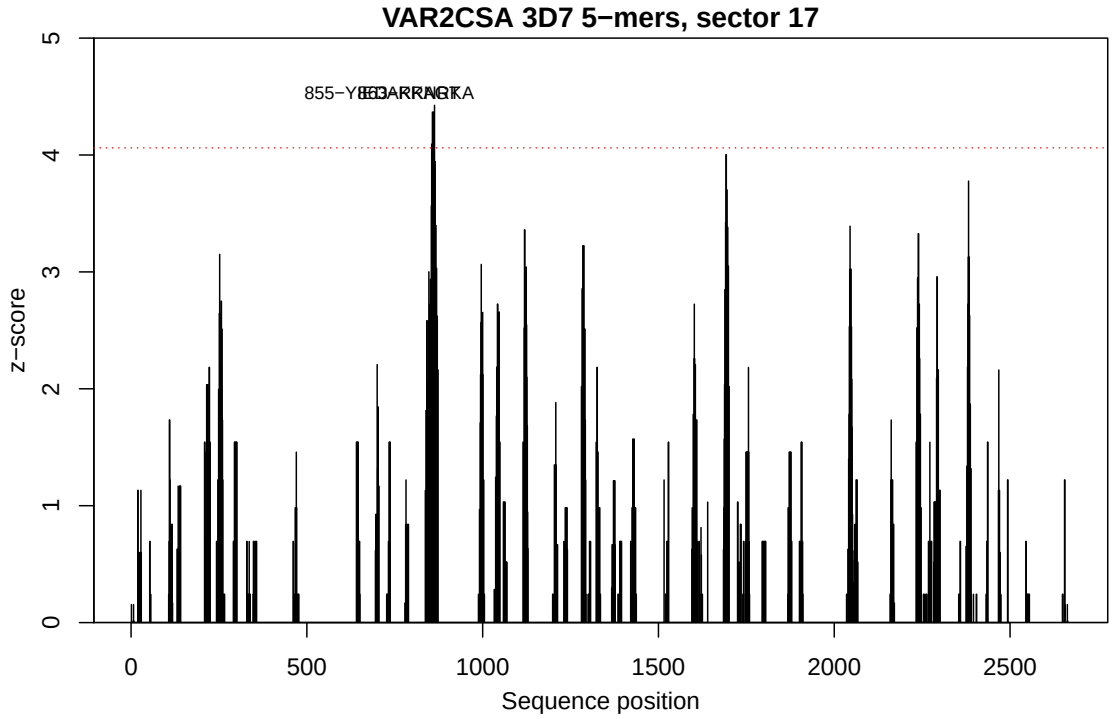

VAR2CSA 3D7 5−mers, sector 12
859−AKRNRKAGT
z−score
0
1
2
3
4
5
6
0
500
1000
1500
2000
2500
Sequence position

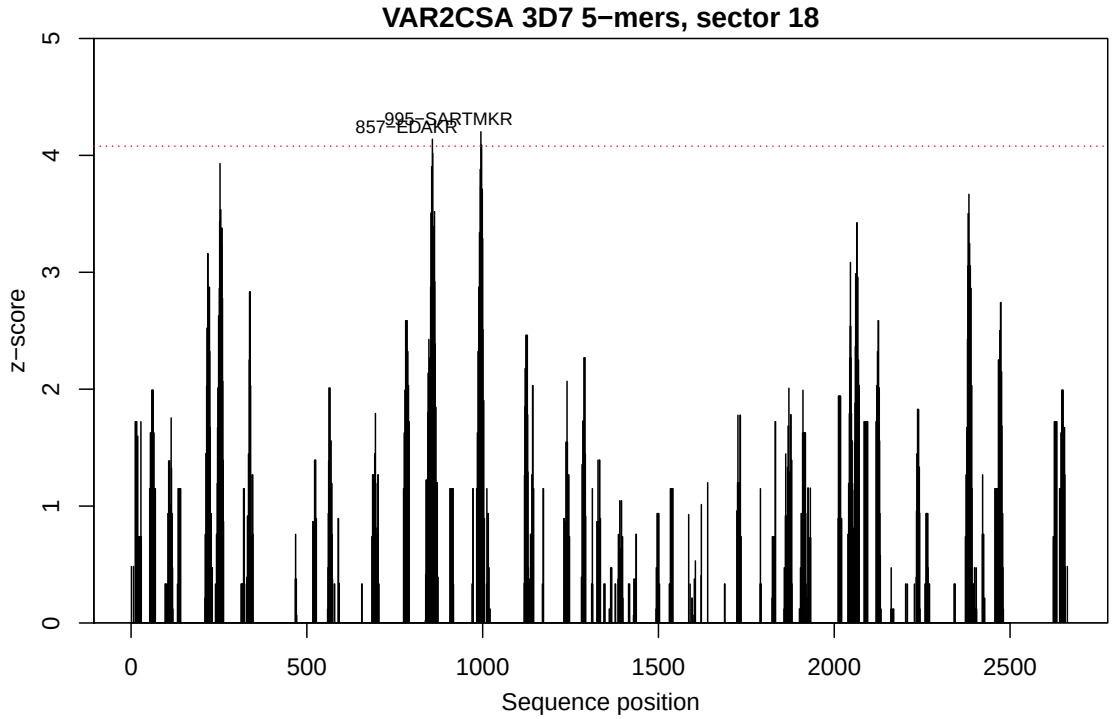

VAR2CSA 3D7 5−mers, sector 13
AGAGA
2653−GAGAGA
1−NHSDS
z−score
Sequence position
0
500
1000
1500
2000
2500
1
2
3
4
5
6
7

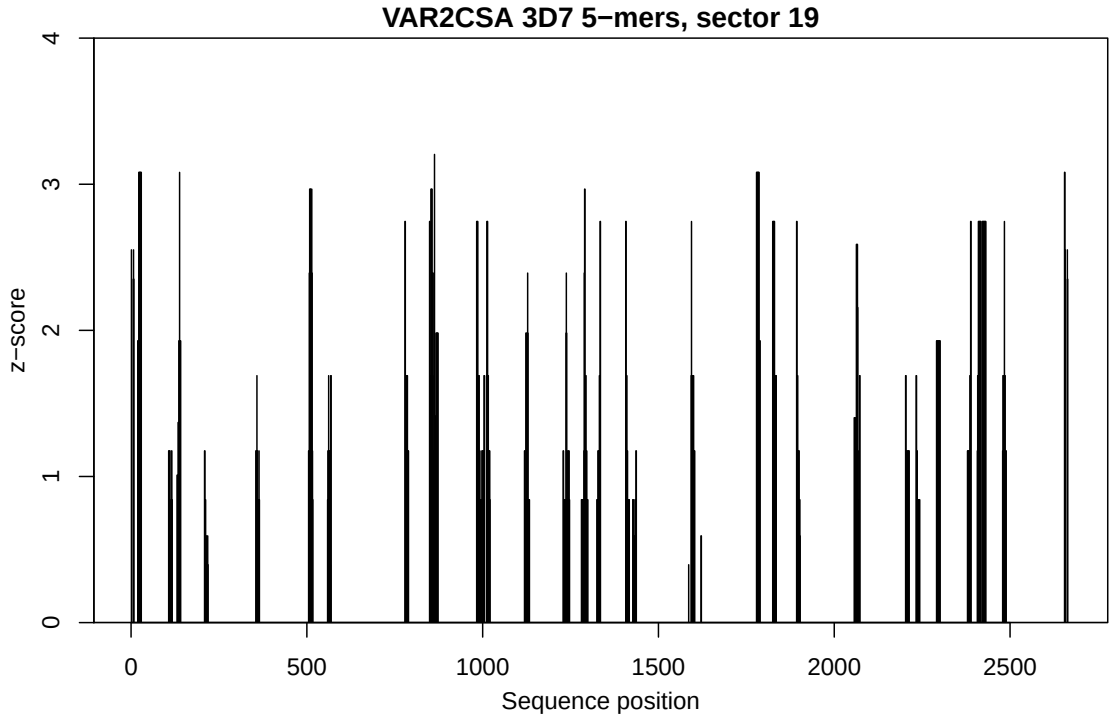

VAR2CSA 3D7 5−mers, sector 14
z−score
Sequence position
0
500
1000
1500
2000
2500
1
2
3
4
5

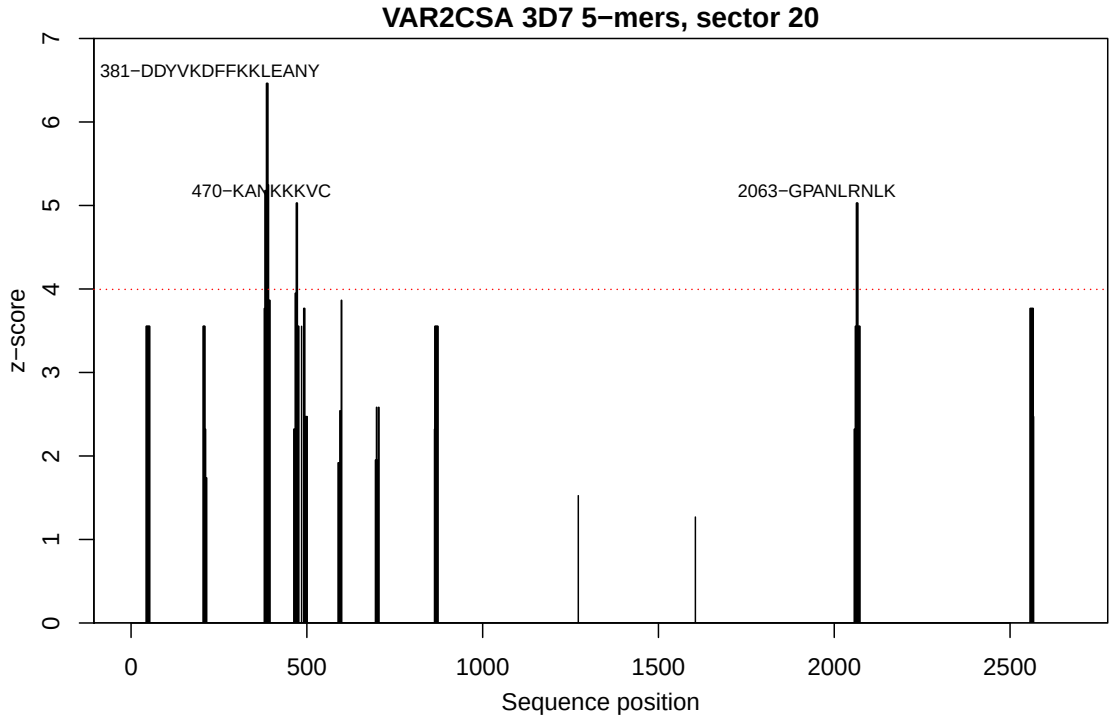

VAR2CSA 3D7 5−mers, sector 15
z−score
0
1
2
3
4
5
0
500
1000
1500
2000
2500
Sequence position

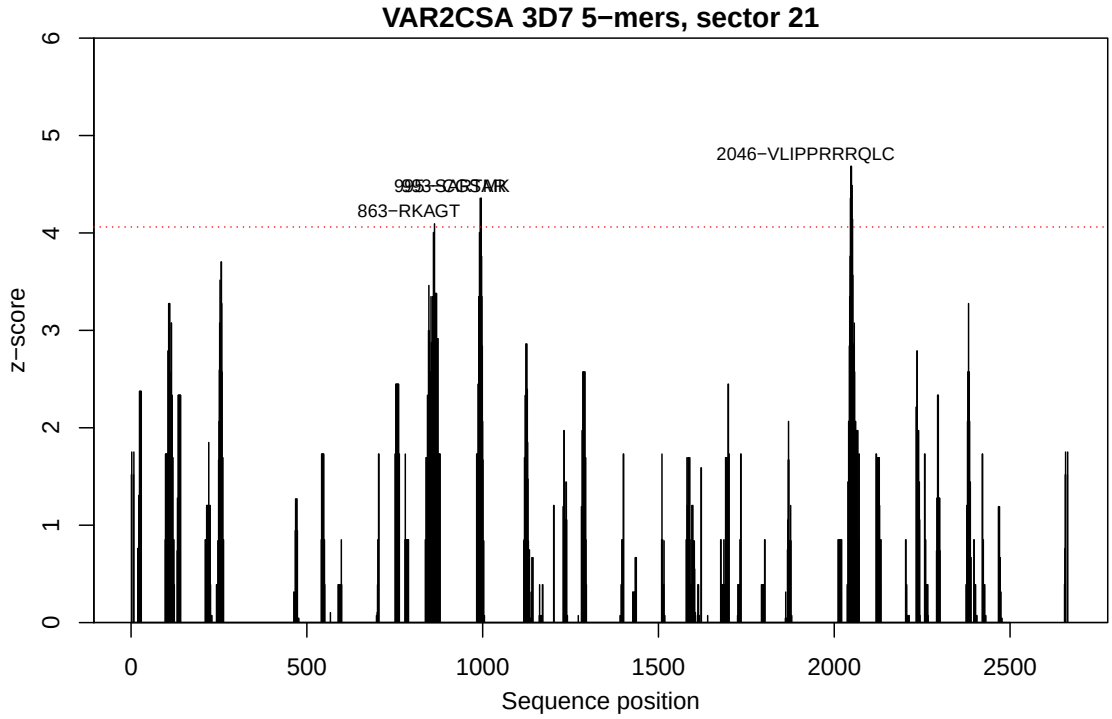

VAR2CSA 3D7 5−mers, sector 16
999−MKRGY
z−score
0
1
2
3
4
5
0
500
1000
1500
2000
2500
Sequence position

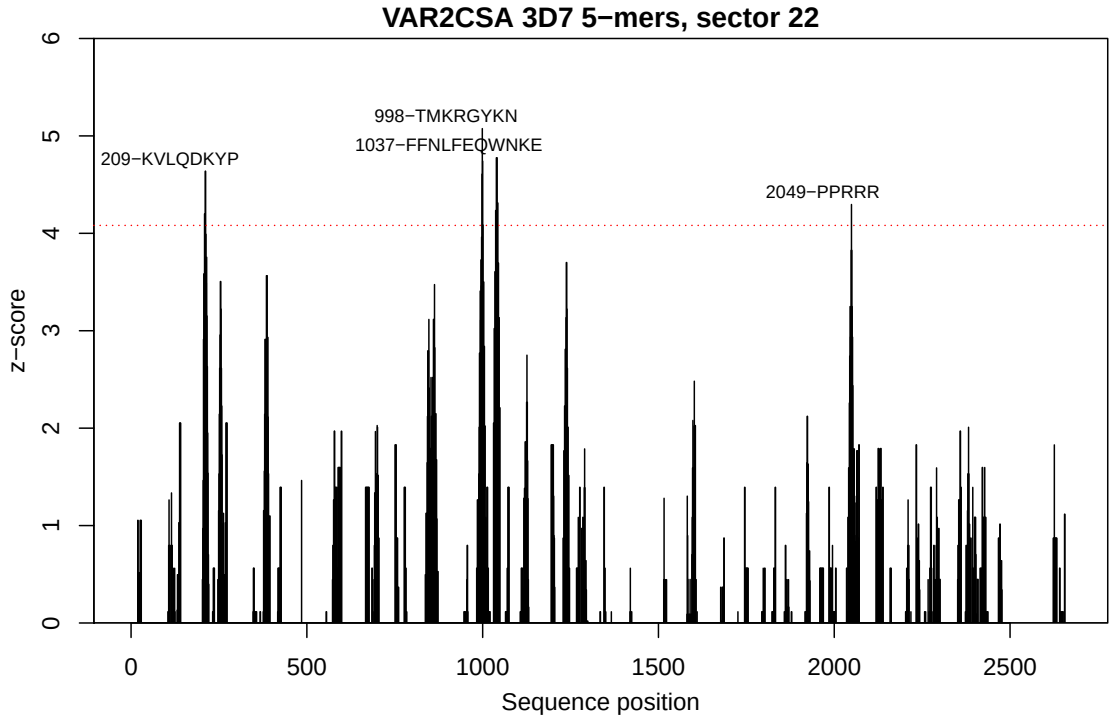

VAR2CSA 3D7 5−mers, sector 17
855−YEDARRAGTKA
z−score
0
1
2
3
4
5
0
500
1000
1500
2000
2500
Sequence position

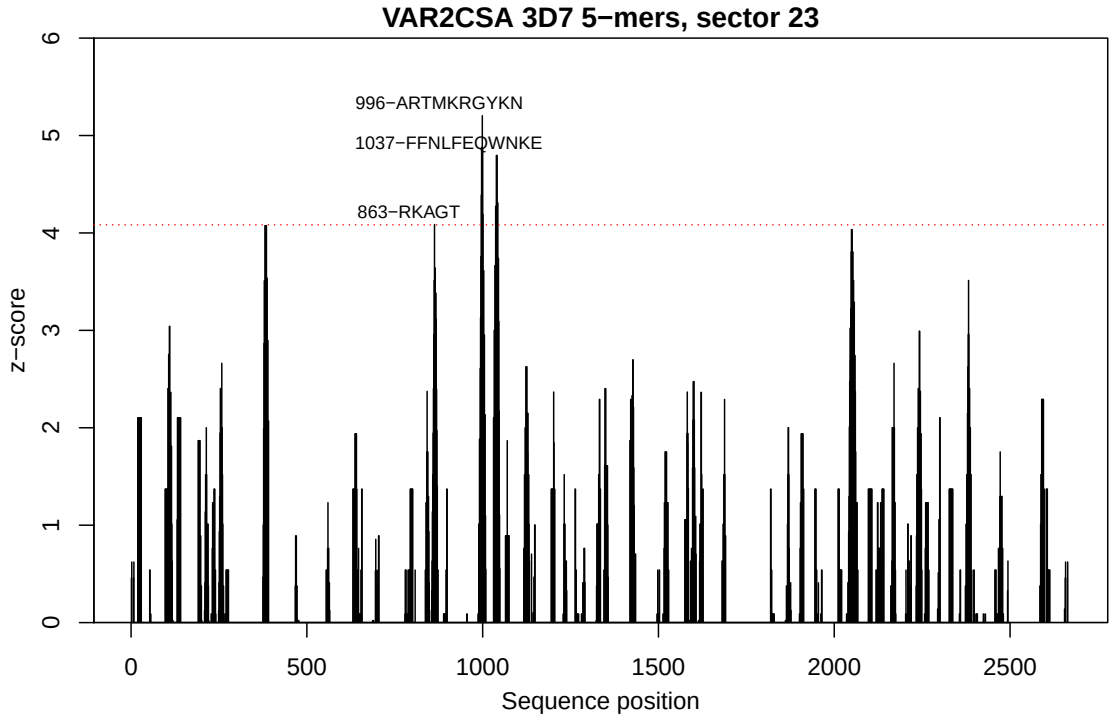

VAR2CSA 3D7 5−mers, sector 18
857−EDAKR
995−SARTMKR
z−score
0
1
2
3
4
5
0
500
1000
1500
2000
2500
Sequence position

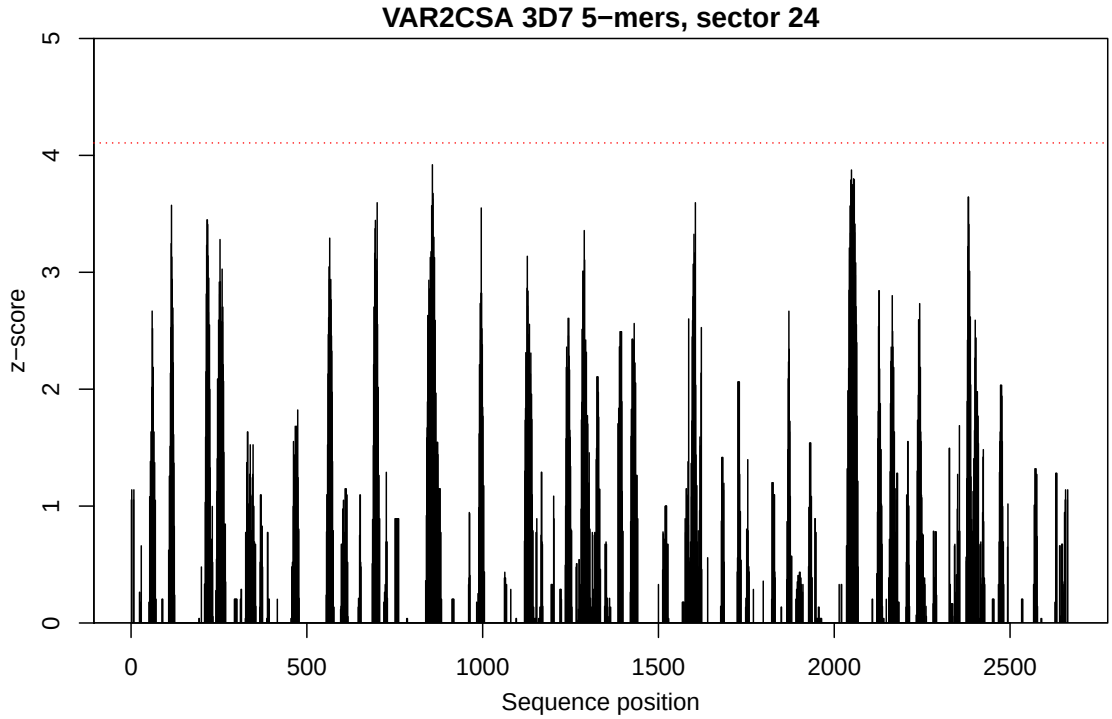

VAR2CSA 3D7 5−mers, sector 19
z−score
0
1
2
3
4
0
500
1000
1500
2000
2500
Sequence position


VAR2CSA 3D7 5−mers, sector 20
381−DDYVKDFFKKLEANY
470−KANKKKVC
2063−GPANLRNLK
z−score
0
1
2
3
4
5
6
7
0
500
1000
1500
2000
2500
Sequence position


VAR2CSA 3D7 5−mers, sector 21
z−score
Sequence position
863−RKAGT
2046−VLIPPRRRQLC
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6


VAR2CSA 3D7 5−mers, sector 22
z−score
Sequence position
209−KVLQDKYP
998−TMKRGYKN
1037−FFNLFEQWNKE
2049−PPRRR
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6


VAR2CSA 3D7 5−mers, sector 23
996−ARTMKRGYKN
1037−FFNLFEQWNKE
863−RKAGT
z−score
Sequence position
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6


VAR2CSA 3D7 5−mers, sector 24
z−score
Sequence position
0
500
1000
1500
2000
2500
0
1
2
3
4
5

Sector	max(z)	max(k-mer)	Start	End	Epitope
1	7.819	VKDFF	386	400	RYDDYVKDFFEKLEA
1	6.909	TMKRG	1008	1024	CGSARTMKRGYKNDNYE
2	5.100	GKSDR	265	271	SGKSDRK
2	4.719	KNLKK	645	653	EGKNLKKRY
2	4.470	EDAKR	868	874	EDAKRNR
2	4.389	YKNIH	2355	2362	YKNIHDRM
2	4.331	RGWRT	260	265	RGWRTS
3	6.142	GWRTS	257	270	LIKRGWRTSGKSDR
3	5.991	MKRGY	1009	1020	GSARTMKRGYKN
3	5.486	LIPPR	2384	2391	GVLIPPRR
3	5.486	LIPPR	2057	2065	GVLIPPRRR
3	5.250	TISKR	2262	2270	QWETISKRY
3	4.876	PNPLL	2681	2689	IPNPLLGLD
3	4.774	EDAKR	867	875	IEDAKRNRK
3	4.383	LKKRY	649	653	LKKRY
3	4.334	IHDRM	2358	2364	IHDRMKK
3	4.160	NYNKF	1138	1143	NYNKFR
3	4.150	SKRWD	853	857	SKRWD
4	7.417	KRGYK	1010	1025	SARTMKRGYKNDNYEL
4	4.951	IGGVG	1411	1420	KIGGVGSSTE
4	4.664	NLFEQ	1054	1061	NLFEQWNK
4	4.607	RTSGK	261	268	GWRTSGKS
4	4.361	GVLIP	2384	2389	GVLIPP
4	4.361	GVLIP	2057	2062	GVLIPP
5	7.080	KRGYK	1010	1024	SARTMKRGYKNDNYE
5	6.920	GKPIP	2671	2689	SGRGELEGKPIPNPLLGLD
5	5.803	KDFFE	389	400	DYVKDFFEKLEA
5	5.532	EDAKR	866	877	HIEDAKRNRKAG
5	4.887	RGWRT	259	267	KRGWRTSGK
5	4.389	IHDRM	2358	2363	IHDRMK
5	4.233	LKKRY	649	653	LKKRY

Table 5.3.4: K-mer mapping of VAR2CSA FCR3 epitopes identified using the HDPMa chip. Sectors 1-5. Absent sectors imply no epitopes were found.

Tables of VAR2CSA 3D7 epitopes identified using the k-mer approach

Sector	max(z)	max(k-mer)	Start	End	Epitope
6	5.538	MKRGY	1011	1020	ARTMKRGYKN
6	4.797	LIPPR	2384	2391	GVLIPPRR
6	4.797	LIPPR	2057	2067	GVLIPPRRRQL
6	4.660	EDAKR	867	876	IEDAKRNRKA
6	4.508	GWRTS	260	266	RGWRTSG
6	4.218	TISKR	2265	2269	TISKR
7	7.260	TMKRG	1008	1024	CGSARTMKRGYKNDNYE
7	6.227	PNPLL	2679	2690	KPIPNPLLGLDS
7	4.255	LIPPR	2386	2390	LIPPR
7	4.255	LIPPR	2059	2063	LIPPR
7	4.168	SGQGD	1512	1516	SGQGD
8	5.573	LIPPR	2384	2391	GVLIPPRR
8	5.573	LIPPR	2053	2073	NKHKGVLIPPRRRQLCFSRIV
8	5.085	KAYGG	2429	2435	KKAYGGA
8	4.735	KRSIQ	2254	2266	EWSRKRSIQWETI
8	4.301	GSART	1009	1015	GSARTMK
8	4.198	TSGKS	264	268	TSGKS
8	4.163	EDAKR	868	873	EDAKRN
9	4.807	KFRSK	1139	1145	YNKFRSK
9	4.725	RTSGK	262	270	WRTSGKSDR
9	4.577	SIQWE	2260	2267	SIQWETIS
9	4.386	VLIPP	2385	2391	VLIPPRR
9	4.386	VLIPP	2058	2064	VLIPPRR
9	4.195	KNLKK	647	651	KNLKK
10	5.997	PLLGL	2680	2692	PIPNPLLGLDSTR
10	5.424	PPRRR	2056	2071	KGVLIPPRRRQLCFSR
10	5.287	RGWRT	258	271	IKRGWRTSGKSDRK
10	5.287	ARTMK	1010	1018	SARTMKRGY
10	5.022	GVLIP	2384	2391	GVLIPPRR
10	4.637	NIHDR	2357	2361	NIHDR
10	4.512	NLKKR	647	652	KNLKKR
10	4.351	EDAKR	868	873	EDAKRN
10	4.139	GARAK	2434	2438	GARAK
10	4.117	KSNKE	1890	1895	KSNKES

Table 5.3.5: K-mer mapping of VAR2CSA FCR3 epitopes identified using the HDPMa chip. Sectors 6-10. Absent sectors imply no epitopes were found.

Sector	max(z)	max(k-mer)	Start	End	Epitope
11	6.275	DLMNE	1955	1969	EEEKGPLDLMNEVLN
11	5.254	YNNAE	1604	1612	KYNNAEKKN
11	4.407	GWRTS	261	266	GWRTSG
11	4.260	KNIHD	2356	2362	KNIHDRM
11	4.257	NAEKK	1218	1222	NAEKK
11	4.256	IPPRR	2060	2064	IPPRR
11	4.182	GVLIP	2384	2388	GVLIP
11	4.182	GVLIP	2057	2061	GVLIP
12	5.132	GSART	1008	1019	CGSARTMKRGYK
12	5.121	PLLGL	2680	2689	PIPNPLLGLD
12	4.693	TISKR	2264	2271	ETISKRYK
12	4.479	NIHDR	2357	2361	NIHDR
12	4.270	GWRTS	261	268	GWRTSGKS
12	4.218	NLKKR	648	652	NLKKR
12	4.180	EDAKR	868	872	EDAKR
13	7.491	LDLMN	1952	1971	TECEEEKGPLDLMNEVLNKM
13	4.346	GWRTS	261	265	GWRTS
13	4.182	TSGKS	264	268	TSGKS
14	5.616	YAVEE	1789	1800	VRYAVEEKNENF
14	5.494	DLMNE	1959	1968	GPLDLMNEVL
14	5.412	RTDWW	1758	1768	ARTDWWENETI
14	5.264	FYDLE	1715	1724	SFYDLEDIIK
14	4.986	DANKK	1701	1707	IDDANKK
14	4.504	GWRTS	261	265	GWRTS
14	4.230	YVPPR	1644	1648	YVPPR
15	5.563	TGKGI	1693	1703	QYNPTGKGIDD
16	6.192	TMKRG	1008	1022	CGSARTMKRGYKNDN
16	5.514	KKLQK	1941	1952	GTVNKKLQKKET
16	4.875	NIHDR	2356	2362	KNIHDRM
16	4.795	GWRTS	261	267	GWRTSGK
16	4.589	GVLIP	2384	2389	GVLIPP
16	4.589	GVLIP	2057	2062	GVLIPP
16	4.223	PLDLM	1960	1964	PLDLM
16	4.223	EEKGP	1956	1960	EEKGP

Table 5.3.6: K-mer mapping of VAR2CSA FCR3 epitopes identified using the HDPMa chip. Sectors 11-16. Absent sectors imply no epitopes were found.

Sector	max(z)	max(k-mer)	Start	End	Epitope
17	5.641	GPLDL	1956	1969	EEKGPLDLMNEVLN
17	4.883	RTMKR	1011	1017	ARTMKRG
17	4.536	GWRTS	261	266	GWRTSG
17	4.507	EDIIK	2125	2130	EDIIKG
17	4.507	EDIIK	1720	1725	EDIIKG
18	5.936	PQNKN	649	664	LKKRYPQNKNSGNKEN
18	5.580	LDLMN	1957	1970	EKGPLDLMNEVLNK
18	5.017	RGWRT	257	269	LIKRGWRTSGKSD
20	8.627	FEKLE	388	404	DDYVKDFFEKLEANYSS
20	5.428	NLKFD	128	139	TYNLENLKFDKI
21	6.154	PLDLM	1955	1968	EEEKGPLDLMNEVL
21	4.633	NLKKR	647	652	KNLKKR
21	4.349	PNPLL	2681	2686	IPNPLL
21	4.193	PPRRR	2061	2065	PPRRR
22	6.038	KDFFE	389	400	DYVKDFFEKLEA
22	5.770	NLFEQ	1049	1063	DVTFFNLFEQWNKEI
22	5.162	KYKDL	366	374	NKYKDLYEQ
22	5.002	YYKYN	1600	1608	VKYYKYNNA
22	4.650	MKRGY	1011	1019	ARTMKRGYK
23	6.589	MKRGY	1008	1023	CGSARTMKRGYKNDNY
23	6.044	SGKSD	261	273	GWRTSGKSDRKKN
23	5.697	NLFEQ	1051	1062	TFFNLFEQWNKE
23	5.425	YEQNK	368	379	YKDLYEQNKNKT
23	4.945	KPIPN	2678	2684	GKPIPNP
23	4.625	DDYVK	388	395	DDYVKDFF
23	4.372	KKRYP	650	654	KKRYP

Table 5.3.7: K-mer mapping of VAR2CSA FCR3 epitopes identified using the HDPMa chip. Sectors 17-23. Absent sectors imply no epitopes were found.

Sector	max(z)	max(k-mer)	Start	End	Epitope
24	6.048	IPPRR	2384	2391	GVLIPPRR
24	6.048	IPPRR	2056	2076	KGVLIPPRRRQLCFSRIVRGP
24	5.305	ARTMK	1009	1018	GSARTMKRGY
24	5.288	KRNRK	867	878	IEDAKRNRKAGT
24	4.962	RGWRT	257	271	LIKRGWRTSGKSDRK
24	4.889	YNKFR	1137	1147	DNYNKFRSKQI
24	4.850	NLKKR	647	655	KNLKKRYPQ
24	4.843	RYKKY	2263	2275	WETISKRYKKYKR
24	4.692	HDRMK	2358	2364	IHDRMKK
24	4.655	GACKR	1519	1528	QGACKRKCEK
24	4.436	GGARA	2428	2438	LKKAYGGARAK

Table 5.3.8: K-mer mapping of VAR2CSA FCR3 epitopes identified using the HDPMa chip. Sector 24. Absent sectors imply no epitopes were found.

Sector	max(z)	max(k-mer)	Start	End	Epitope
1	6.066	DDYVK	378	391	SRYDDYVKDFFKKL
1	4.140	GYKND	1002	1007	GYKNDN
2	5.040	EDAKR	853	863	SKYIEDAKRNR
2	4.657	RGWRT	252	261	KRGWRTSGKS
2	4.139	KNTFK	2268	2273	KNTFKN
2	4.139	EFKNT	2266	2270	EFKNT
2	4.062	FQRKQ	1127	1131	FQRKQ
3	4.341	GVLIP	2045	2049	GVLIP
3	4.240	ARTMK	996	1000	ARTMK
3	4.128	LIPPR	2047	2051	LIPPR
4	5.435	MKRGY	997	1005	RTMKRGYKN
4	4.967	NLFEQ	1037	1045	FFNLFEQWN
5	5.426	MKRGY	994	1005	GSARTMKRGYKN
5	4.419	DDYVK	381	386	DDYVKD
5	4.189	KRNRK	860	864	KRNRK
6	4.239	MKRGY	999	1005	MKRGYKN
7	5.590	MKRGY	995	1009	SARTMKRGYKNDNYE
7	4.728	IKRGW	249	259	LLIKRGWRTSG
8	4.415	LIPPR	2046	2053	VLIPPRRR
8	4.319	NKFQR	1124	1130	YNKFQRK
9	4.513	GAGAG	7	12	GAGAGA
9	4.513	GAGAG	2663	2668	GAGAGA
9	4.513	GAGAG	2657	2662	GAGAGA
9	4.513	GAGAG	1	6	GAGAGA
12	4.684	RNRKA	859	867	AKRNRKAGT
13	6.157	AGAGA	7	12	GAGAGA
13	6.157	AGAGA	2663	2668	GAGAGA
13	6.157	AGAGA	2657	2662	GAGAGA
13	6.157	AGAGA	1	6	GAGAGA
13	4.611	NHSDS	71	75	NHSDS
16	4.305	LIPPR	2047	2051	LIPPR
16	4.305	GVLIP	2045	2049	GVLIP
16	4.074	MKRGY	999	1003	MKRGY
17	4.367	EDAKR	855	865	YIEDAKRNRKA
18	4.202	SARTM	995	1001	SARTMKR
18	4.138	EDAKR	857	861	EDAKR

Table 5.3.9: K-mer mapping of VAR2CSA 3D7 epitopes identified using the HDPMa chip. Sectors 1-18. Absent sectors imply no epitopes were found.

Sector	max(z)	max(k-mer)	Start	End	Epitope
20	6.462	KDFFK	381	395	DDYVKDFFKKLEANY
20	5.028	KANKK	470	477	KANKKKVC
20	5.028	GPANL	2063	2071	GPANLRNLK
21	4.684	LIPPR	2046	2056	VLIPPRRRQLC
21	4.356	SARTM	995	1000	SARTMK
21	4.356	CGSAR	993	997	CGSAR
21	4.094	RKAGT	863	867	RKAGT
22	5.075	MKRGY	998	1005	TMKRGYKN
22	4.775	NLFEQ	1037	1047	FFNLFEQWNKE
22	4.637	LQDKY	209	216	KVLQDKYP
23	5.204	MKRGY	996	1005	ARTMKRGYKN
23	4.795	NLFEQ	1037	1047	FFNLFEQWNKE
23	4.086	RKAGT	863	867	RKAGT

Table 5.3.10: K-mer mapping of VAR2CSA 3D7 epitopes identified using the HDPMa chip. Sectors 20-23. Absent sectors imply no epitopes were found.

5.3.3 FCR3 and 3D7 sequences

>VAR2CSA_FCR3 2715 residues
GAGAGAGAGAGAGAGGYLLEFMDSTSTIANKIEEYLGAKSDDSKIDELLKADPSEVEYYRSGGDGDYLKNNICKITVNHS
DSGKYDPCEKKLPPYDDNDQWKCQQNSSDGSGKPENICVPPRRERLCTYNLENLKFDKIRDNNAFLADVLLTARNEGEKI
VQNHPDTNSSNVCNALERSFADLADIIRGTDQWKGTNSNLEKNLKQMFAKIRENDKVLQDKYPKDQKYTKLREAWWNANR
QKVWEVITCGARSNDLLIKRGWRTSGKSDRKKNFELCRKCGHYEKEVPTKLDYVPQFLRWLTEWIEDFYREKQNLIDDME
RHREECTREDHKSKEGTSYCSTCKDKCKKYCECVKKWKTEWENQENKYKDLYEQNKNKTSQKNTSRYDDYVKDFFEKLEA
NYSSLENYIKGDPYFAEYATKLSFILNPSDANNPSGETANHNDEACNCNESGISSVGQAQTSGPSSNKTCITHSSIKTNK
KKECKDVKLGVRENDKDLKICVIEDTSLSGVDNCCCQDLLGILQENCSDNKRGSSSNDSCDNKNQDECQKKLEKVFASLT
NGYKCDKCKSGTSRSKKKWIWKKSSGNEEGLQEEYANTIGLPPRTQSLYLGNLPKLENVCEDVKDINFDTKEKFLAGCLI
VSFHEGKNLKKRYPQNKNSGNKENLCKALEYSFADYGDLIKGTSIWDNEYTKDLELNLQNNFGKLFGKYIKKNNTAEQDT
SYSSLDELRESWWNTNKKYIWTAMKHGAEMNITTCNADGSVTGSGSSCDDIPTIDLIPQYLRFLQEWVENFCEQRQAKVK
DVITNCKSCKESGNKCKTECKTKCKDECEKYKKFIEACGTAGGGIGTAGSPWSKRWDQIYKRYSKHIEDAKRNRKAGTKN
CGTSSTTNAAASTDENKCVQSDIDSFFKHLIDIGLTTPSSYLSNVLDDNICGADKAPWTTYTTYTTTEKCNKERDKSKSQ
SSDTLVVVNVPSPLGNTPYRYKYACQCKIPTNEETCDDRKEYMNQWSCGSARTMKRGYKNDNYELCKYNGVDVKPTTVRS
NSSKLDGNDVTFFNLFEQWNKEIQYQIEQYMTNANISCIDEKEVLDSVSDEGTPKVRGGYEDGRNNNTDQGTNCKEKCKC
YKLWIEKINDQWGKQKDNYNKFRSKQIYDANKGSQNKKVVSLSNFLFFSCWEEYIQKYFNGDWSKIKNIGSDTFEFLIKK
CGNNSAHGEEIFSEKLKNAEKKCKENESTDTNINKSETSCDLNATNYIRGCQSKTYDGKIFPGKGGEKQWICKDTIIHGD
TNGACIPPRTQNLCVGELWDKSYGGRSNIKNDTKELLKEKIKNAIHKETELLYEYHDTGTAIISKNDKKGQKGKNDPNGL
PKGFCHAVQRSFIDYKNMILGTSVNIYEHIGKLQEDIKKIIEKGTPQQKDKIGGVGSSTENVNAWWKGIEREMWDAVRCA
ITKINKKNNNSIFNGDECGVSPPTGNDEDQSVSWFKEWGEQFCIERLRYEQNIREACTINGKNEKKCINSKSGQGDKIQG
ACKRKCEKYKKYISEKKQEWDKQKTKYENKYVGKSASDLLKENYPECISANFDFIFNDNIEYKTYYPYGDYSSICSCEQV
KYYKYNNAEKKNNKSLCYEKDNDMTWSKKYIKKLENGRSLEGVYVPPRRQQLCLYELFPIIIKNEEGMEKAKEELLETLQ
IVAEREAYYLWKQYNPTGKGIDDANKKACCAIRGSFYDLEDIIKGNDLVHDEYTKYIDSKLNEIFGSSNTNDIDTKRART
DWWENETITNGTDRKTIRQLVWDAMQSGVRYAVEEKNENFPLCMGVEHIGIAKPQFIRWLEEWTNEFCEKYTKYFEDMKS
KCDPPKRADTCGDNSNIECKKACANYTNWLNPKRIEWNGMSNYYNKIYRKSNKESEDGKDYSMIMAPTVIDYLNKRCHGE
INGNYICCSCKNIGAYNTTSGTVNKKLQKKETECEEEKGPLDLMNEVLNKMDKKYSAHKMKCTEVYLEHVEEQLNEIDNA
IKDYKLYPLDRCFDDQTKMKVCDLIADAIGCKDKTKLDELDEWNDMDLRGTYNKHKGVLIPPRRRQLCFSRIVRGPANLR
SLNEFKEEILKGAQSEGKFLGNYYKEHKDKEKALEAMKNSFYDYEDIIKGTDMLTNIEFKDIKIKLDRLLEKETNNTKKA
EDWWKTNKKSIWNAMLCGYKKSGNKIIDPSWCTIPTTETPPQFLRWIKEWGTNVCIQKQEHKEYVKSKCSNVTNLGAQAS
ESNNCTSEIKKYQEWSRKRSIQWETISKRYKKYKRMDILKDVKEPDANTYLREHCSKCPCGFNDMEEMNNNEDNEKEAFK
QIKEQVKIPAELEDVIYRIKHHEYDKGNDYICNKYKNIHDRMKKNNGNFVTDNFVKKSWEISNGVLIPPRRKNLFLYIDP
SKICEYKKDPKLFKDFIYWSAFTEVERLKKAYGGARAKVVHAMKYSFTDIGSIIKGDDMMEKNSSDKIGKILGDTDGQNE
KRKKWWDMNKYHIWESMLCGYREAEGDTETNENCRFPDIESVPQFLRWFQEWSENFCDRRQKLYDKLNSECISAECTNGS
VDNSKCTHACVNYKNYILTKKTEYEIQTNKYDNEFKNKNSNDKDAPDYLKEKCNDNKCECLNKHIDDKNKTWKNPYETLE
DTFKSKCDCPKPLPSPIKPDDLPPQADEPFSGRGELEGKPIPNPLLGLDSTRTGHHHHHHGAGAGAGAGAGAGAG
>VAR2CSA_3D7 2671 residues
GAGAGAGAGAGAGAGMDKSSIANKIEAYLGAKSDDSKIDQSLKADPSEVQYYGSGGDGYYLRKNICKITVNHSDSGTNDP
CDRIPPPYGDNDQWKCAIILSKVSEKPENVFVPPRRQRMCINNLEKLNVDKIRDKHAFLADVLLTARNEGERIVQNHPDT
NSSNVCNALERSFADIADIIRGTDLWKGTNSNLEQNLKQMFAKIRENDKVLQDKYPKDQNYRKLREDWWNANRQKVWEVI
TCGARSNDLLIKRGWRTSGKSNGDNKLELCRKCGHYEEKVPTKLDYVPQFLRWLTEWIEDFYREKQNLIDDMERHREECT
SEDHKSKEGTSYCSTCKDKCKKYCECVKKWKSEWENQKNKYTELYQQNKNETSQKNTSRYDDYVKDFFKKLEANYSSLEN
YIKGDPYFAEYATKLSFILNSSDANNPSEKIQKNNDEVCNCNESGIASVEQEQISDPSSNKTCITHSSIKANKKKVCKHV
KLGVRENDKDLRVCVIEHTSLSGVENCCCQDFLRILQENCSDNKSGSSSNGSCNNKNQEACEKNLEKVLASLTNCYKCDK
CKSEQSKKNNKNWIWKKSSGKEGGLQKEYANTIGLPPRTQSLCLVVCLDEKGKKTQELKNIRTNSELLKEWIIAAFHEGK
NLKPSHEKKNDDNGKKLCKALEYSFADYGDLIKGTSIWDNEYTKDLELNLQKIFGKLFRKYIKKNNTAEQDTSYSSLDEL
RESWWNTNKKYIWLAMKHGAGMNSTTCCGDGSVTGSGSSCDDIPTIDLIPQYLRFLQEWVEHFCKQRQEKVKPVIENCKS
CKESGGTCNGECKTECKNKCEVYKKFIEDCKGGDGTAGSSWVKRWDQIYKRYSKYIEDAKRNRKAGTKNCGPSSTTNAAE
NKCVQSDIDSFFKHLIDIGLTTPSSYLSIVLDDNICGADKAPWTTYTTYTTTEKCNKETDKSKLQQCNTAVVVNVPSPLG
NTPHGYKYACQCKIPTNEETCDDRKEYMNQWSCGSARTMKRGYKNDNYELCKYNGVDVKPTTVRSNSSKLDDKDVTFFNL
FEQWNKEIQYQIEQYMTNTKISCNNEKNVLSRVSDEAAQPKFSDNERDRNSITHEDKNCKEKCKCYSLWIEKINDQWDKQ
KDNYNKFQRKQIYDANKGSQNKKVVSLSNFLFFSCWEEYIQKYFNGDWSKIKNIGSDTFEFLIKKCGNDSGDGETIFSEK
LNNAEKKCKENESTNNKMKSSETSCDCSEPIYIRGCQPKIYDGKIFPGKGGEKQWICKDTIIHGDTNGACIPPRTQNLCV
GELWDKRYGGRSNIKNDTKESLKQKIKNAIQKETELLYEYHDKGTAIISRNPMKGQKEKEEKNNDSNGLPKGFCHAVQRS
FIDYKNMILGTSVNIYEYIGKLQEDIKKIIEKGTTKQNGKTVGSGAENVNAWWKGIEGEMWDAVRCAITKINKKQKKNGT
FSIDECGIFPPTGNDEDQSVSWFKEWSEQFCIERLQYEKNIRDACTNNGQGDKIQGDCKRKCEEYKKYISEKKQEWDKQK

TKYENKYVGKSASDLLKENYPECISANFDFIFNDNIEYKTYYPYGDYSSICSCEQVKYYEYNNAEKKNNKSLCHEKGNDR
TWSKKYIKKLENGRTLEGVYVPPRRQQLCLYELFPIIIKNKNDITNAKKELLETLQIVAEREAYYLWKQYHAHNDTTYLA
HKKACCAIRGSFYDLEDIIKGNDLVHDEYTKYIDSKLNEIFDSSNKNDIETKRARTDWWENEAIAVPNITGANKSDPKTI
RQLVWDAMQSGVRKAIDEEKEKKKPNENFPPCMGVQHIGIAKPQFIRWLEEWTNEFCEKYTKYFEDMKSNCNLRKGADDC
DDNSNIECKKACANYTNWLNPKRIEWNGMSNYYNKIYRKSNKESEDGKDYSMIMEPTVIDYLNKRCNGEINGNYICCSCK
NIGENSTSGTVNKKLQKKETQCEDNKGPLDLMNKVLNKMDPKYSEHKMKCTEVYLEHVEEQLKEIDNAIKDYKLYPLDRC
FDDKSKMKVCDLIGDAIGCKHKTKLDELDEWNDVDMRDPYNKYKGVLIPPRRRQLCFSRIVRGPANLRNLKEFKEEILKG
AQSEGKFLGNYYNEDKDKEKALEAMKNSFYDYEYIIKGSDMLTNIQFKDIKRKLDRLLEKETNNTEKVDDWWETNKKSIW
NAMLCGYKKSGNKIIDPSWCTIPTTETPPQFLRWIKEWGTNVCIQKEEHKEYVKSKCSNVTNLGAQESESKNCTSEIKKY
QEWSRKRSIQWEAISEGYKKYKGMDEFKNTFKNIKEPDANEPNANEYLKKHCSKCPCGFNDMQEITKYTNIGNEAFKQIK
EQVDIPAELEDVIYRLKHHEYDKGNDYICNKYKNINVNMKKNNDDTWTDLVKNSSDINKGVLLPPRRKNLFLKIDESDIC
KYKRDPKLFKDFIYSSAISEVERLKKVYGEAKTKVVHAMKYSFADIGSIIKGDDMMENNSSDKIGKILGDGVGQNEKRKK
WWDMNKYHIWESMLCGYKHAYGNISENDRKMLDIPNNDDEHQFLRWFQEWTENFCTKRNELYENMVTACNSAKCNTSNGS
VDKKECTEACKNYSNFILIKKKEYQSLNSQYDMNYKETKAEKKESPEYFKDKCNGECSCLSEYFKDETRWKNPYETLDDT
EVKNNCMCKPPPPASNGAGAGAGAGAGAGAG

5.4 Part V - Development and Application of Diversity Covering Sequence Generator

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. A template, which can be used to format a PhD thesis with this look and feel, has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/ or from the author at suchow@post.harvard.edu.